Species Distribution Modelling of Corals and Sponges in the Maritimes Region for Use in the Identification of Significant Benthic Areas

L. Beazley, E. Kenchington, F.J. Murillo, C. Lirette, J. Guijarro, A. McMillan, A. Knudby

Ocean and Ecosystem Sciences Division
Maritimes Region
Fisheries and Oceans Canada

Bedford Institute of Oceanography
PO Box 1006
Dartmouth, Nova Scotia
Canada B2Y 4A2

2016

Canadian Technical Report of Fisheries and Aquatic Sciences 3172

Fisheries and Oceans Canada | Pêches et Océans Canada

Canada

Canadian Technical Report of Fisheries and Aquatic Sciences

Technical reports contain scientific and technical information that contributes to existing knowledge but which is not normally appropriate for primary literature. Technical reports are directed primarily toward a worldwide audience and have an international distribution. No restriction is placed on subject matter and the series reflects the broad interests and policies of Fisheries and Oceans Canada, namely, fisheries and aquatic sciences.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts.*

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. Numbers 715-924 were issued as Department of Fisheries and Environment, Fisheries and Marine Service Technical Reports. The current series name was changed with report number 925.

Rapport technique canadien des sciences halieutiques et aquatiques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles, mais qui ne sont pas normalement appropriés pour la publication dans un journal scientifique. Les rapports techniques sont destinés essentiellement à un public international et ils sont distribués à cet échelon. Il n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques de Pêches et Océans Canada, c'est-à-dire les sciences halieutiques et aquatiques.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact Fig. au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques.*

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom Fig. sur la couverture et la page du titre.

Les numéros 1 à 456 de cette série ont été publiés à titre de Rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457 à 714 sont parus à titre de Rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Les numéros 715 à 924 ont été publiés à titre de Rapports techniques du Service des pêches et de la mer, ministère des Pêches et de l'Environnement. Le nom actuel de la série a été établi lors de la parution du numéro 925.

Canadian Technical Report of
Fisheries and Aquatic Sciences 3172

2016

Species Distribution Modelling of Corals and Sponges in the Maritimes Region for Use in the Identification of Significant Benthic Areas

by

L. Beazley[1], E. Kenchington[1], F.J. Murillo[1], C. Lirette[1], J. Guijarro[1], A. McMillan[2], A. Knudby[3]

[1]Ocean and Ecosystem Sciences Division
Maritimes Region
Fisheries and Oceans Canada
Bedford Institute of Oceanography
P.O. Box 1006
Dartmouth, N.S.
Canada B2Y 4A2

[2]Department of Geography
Simon Fraser University
Burnaby, B.C.
Canada V5A 1S6

[3]Department of Geography, Environment and Geomatics
University of Ottawa
Ottawa, O.N.
Canada K1N 6N5

© Her Majesty the Queen in Right of Canada, 2016.
Cat. No. Fs97-6/3172E-PDF ISBN 978-0-660-05696-8 ISSN 1488-5379

Correct citation for this publication:

Beazley, L., Kenchington, E., Murillo, F.J., Lirette, C., Guijarro, J., McMillan, A., and Knudby, A. 2016. Species Distribution Modelling of Corals and Sponges in the Maritimes Region for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3172: vi + 189p.

TABLE OF CONTENTS

ABSTRACT v
RÉSUMÉ vi
INTRODUCTION 1
MATERIALS AND METHODS 3
Study Area 3
Environmental Data 4
Response Data 4
Random Forest Modelling 10
Model Evaluation 11
Presence-Absence Response Data – Classification Model 11
Biomass Response Data – Regression Model 13
Model Extrapolation 13
Ecological Interpretation 14
Alternative Prediction Models 15
RESULTS 16
Sponges (Porifera) 16
Data Sources and Distribution 16
Model 1 - Balanced Species Prevalence 19
Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence 26
Model Selection 33
Validation of Selected Model Using Independent Data 33
Prediction of Sponge Biomass Using Random Forest 35
Vazella pourtalesi (Russian Hat sponge) 41
Data Sources and Distribution 41
Model 1 - Balanced Species Prevalence 43
Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence 50
Model 3 - Addition of Commercial Records and *In Situ* Benthic Imagery Observations 56
Model Selection 64
Prediction of *Vazella pourtalesi* Biomass Using Random Forest 65
Sea Pens (Pennatulacea) 70
Data Sources and Distribution 70
Model 1 - Balanced Species Prevalence 72
Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence 79
Model 3 - Addition of *In Situ* Benthic Imagery Observations 85
Model Selection 93
Validation of Selected Model Using Independent Data 93
Prediction of Sea Pen Biomass Using Random Forest 95
Large Gorgonian Corals 100
Data Sources and Distribution 100
Model 1 - Balanced Species Prevalence 102
Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence 109
Model 3 - Addition of *In Situ* Benthic Imagery Observations 115
Model Selection 123

Validation of Selected Model Using Independent Data 123
Prediction of Large Gorgonian Coral Biomass Using Random Forest 125
Small Gorgonian Corals 130
Data Sources and Distribution 130
Model 1 - Balanced Species Prevalence 132
Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence 139
Model 3 - Addition of *In Situ* Benthic Imagery Observations 144
Model Selection 152
Validation of Selected Model Using Independent Data 152
Prediction of Small Gorgonian Coral Biomass Using Random Forest 154
DISCUSSION 159
ACKNOWLEDGMENTS 163
REFERENCES 164
APPENDIX 1- Alternative Prediction Models- Generalized Additive Models for Predicting Coral and Sponge Biomass in the Maritimes Region 168
Sponges (Porifera) 169
Vazella pourtalesi (Russian Hat Sponge) 174
Sea Pens (Pennatulacea) 179
Large Gorgonian Corals 182
Small Gorgonian Corals 185

ABSTRACT

Beazley, L., Kenchington, E., Murillo, F.J., Lirette, C., Guijarro, J., McMillan, A., and Knudby, A. 2016. Species Distribution Modelling of Corals and Sponges in the Maritimes Region for Use in the Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3172: vi + 189p.

Effective fisheries and habitat management processes require knowledge of the distribution of areas of high ecological or biological significance. On the Scotian Shelf and Slope, a number of benthic ecologically or biologically significant areas consisting of habitat-forming species such as sponges and deep-water corals have been identified. However, knowledge of their spatial distribution is largely based on targeted surveys that are limited in their spatial extent. We used a species distribution modelling approach called random forest (RF) to predict the probability of occurrence and biomass of sponges, sea pens, and large and small gorgonian corals across the entire spatial extent of Fisheries and Oceans Canada's (DFO) Maritimes Region. We also modelled the rare sponge *Vazella pourtalesi*, which forms the largest known aggregation of its kind on the Scotian Shelf. We utilized a number of data sources including DFO multispecies trawl catch data and *in situ* benthic imagery observations. Most models had excellent predictive capacity with cross-validated Area Under the Receiver Operating Characteristic Curve (AUC) values ranging from 0.760 to 0.977. Areas of suitable habitat were identified for each taxon and were contrasted against their known distribution and when applicable, the location of closure areas designated for their protection. Generalized additive models (GAMs) were developed to predict the biomass distribution of each taxonomic group and serve as a comparison to the RF models. The RF and GAM models provided comparable results, although GAMs provided superior predictions of biomass along the continental slope for some taxonomic groups. In the absence of data observations, the results of this study could be used to identify the potential distribution of sensitive benthic taxa for use in fisheries and habitat management applications. These results could also be used to refine significant concentrations of these taxa as identified through the kernel density analyses.

RÉSUMÉ

Beazley, L., Kenchington, E., Murillo, F.J., Lirette, C., Guijarro, J., McMillan, A., et Knudby, A. 2016. Modélisation de la répartition des espèces de coraux et d'éponges dans la région des Maritimes aux fins d'utilisation dans la détermination des zones benthiques d'importance. Can. Tech. Rep. Fish. Aquat. 3172: vi + 189p.

Pour être efficaces, les processus de gestion des pêches et de l'habitat exigent la connaissance de la répartition des zones de grande importance sur le plan écologique ou biologique. Sur le plateau et le talus néo-écossais, un certain nombre de zones benthiques d'importance écologique ou biologique composées d'espèces faisant office d'habitat, telles que les éponges et les coraux en eau profonde, ont été désignées. Toutefois, la connaissance de leur répartition spatiale se fonde en grande partie sur des relevés ciblés dont l'étendue spatiale est limitée. Nous nous sommes servis d'une approche de modélisation de la répartition des espèces appelée modèle de forêts aléatoires (RF) pour prévoir la probabilité de la présence et la biomasse des éponges, des pennatules et des grandes et petites gorgones dans l'ensemble de l'étendue spatiale de la Région des Maritimes de Pêches et Océans Canada (MPO). Nous avons aussi modélisé *Vazella pourtalesi*, une éponge rare qui forme la plus grande concentration connue du genre sur le plateau néo-écossais. Nous avons utilisé un certain nombre de sources de données, y compris les données sur les prises des relevés plurispécifiques au chalut du MPO et celles provenant des observations *in situ* par imagerie benthique. La plupart des modèles avaient une excellente efficacité de prévision selon des valeurs contre-validées de l'aire sous la courbe de la fonction d'efficacité du récepteur variant de 0,760 à 0,977. Les zones d'habitat propice ont été désignées pour chaque taxon et ont été comparées à l'aire de répartition de l'espèce et, le cas échéant, aux emplacements des zones de fermeture visant sa protection. Des modèles additifs généralisés ont été élaborés pour prédire la répartition de la biomasse de chaque groupe taxonomique et servent de points de comparaison aux modèles RF. Les résultats obtenus par les modèles RF et les modèles additifs généralisés étaient similaires, cependant les prévisions de la biomasse par les modèles additifs généralisés étaient meilleures pour certains groupes taxonomiques le long de la pente continentale. En l'absence de données d'observation, les résultats de la présente étude pourraient servir à déterminer l'aire de répartition potentielle des taxons benthiques vulnérables aux fins d'utilisation dans les applications de gestion des pêches et de l'habitat. Ces résultats pourraient aussi être utilisés pour mieux définir les concentrations importantes de ces taxons repérées dans le cadre des analyses de noyaux de densité.

INTRODUCTION

The Scotian Shelf is a wide, submerged portion of the continental shelf situated off Nova Scotia. It reaches 700 km in length and is up to 230 km wide from the coast to the shelf edge. It is separated from the Gulf of Maine by the Northeast Channel in the southwest, and from the Newfoundland Shelf by the Laurentian Channel in the east. The Scotian Shelf is characterized by a number of valleys, ridges, shallow banks, and deep basins that support a rich diversity of habitats and species, including several commercially-important fishes and invertebrates (Drinkwater et al., 2002). The Scotian Slope and the deep canyons that indent it support a high diversity of sensitive benthic invertebrates such as corals and sponges (Mortensen et al., 2006; Gordon and Kenchington, 2007; Cogswell et al., 2009). In 2014, Fisheries and Oceans Canada (DFO) identified eighteen Ecologically or Biologically Significant Areas (EBSAs) in the offshore component of the Scotian Shelf Biogeographic Region (DFO, 2014). Seventeen of these EBSAs occurred on the Scotian Shelf or Slope, while one was identified in the deeper waters beyond the slope. A number of different ecological or biological data layers were considered to help evaluate and identify these EBSAs, including, areas of high biological productivity or biomass, high fish and invertebrate diversity, important habitats for fishes and invertebrates, coral and sponge occurrences, and critical habitat for species at risk. Although the delineation of an EBSA does not impart immediate conservation status, it does draw attention to an area that has particularly high ecological or biological significance, information that will be useful during broader oceans planning and management processes including marine protected area (MPA) network design (DFO, 2004; 2014).

Kenchington (2014) compiled information on marine benthic species and habitats that are recognized in other jurisdictions as meeting EBSA or similar criteria. Fourteen structure-forming, biogenic habitats such as kelp forests and sponge aggregations that are known or are likely to occur in the Maritimes Region were identified and ranked against DFO 2004 EBSA criteria and additional EBSA criteria from the Convention on Biological Diversity (CBD, 2009). Although some of these benthic EBSAs, such as large deep-water corals, have already received considerable attention in the Maritimes Region, knowledge of their spatial distribution is largely based on targeted surveys that are limited in their spatial extent.

Statistical tools like species distribution modelling (SDM), that are used to predict the distribution of a species in unsampled areas based on its species-environment relationship in sampled areas, are becoming more widely considered in fisheries and habitat management processes. SDM has particular relevance to the identification of benthic EBSAs through its ability to extrapolate predictions of species' occurrence to data-poor areas. Species distribution modelling of sensitive benthic fauna in Atlantic Canada has largely been limited to certain deep-water taxa (see Bryan and Metaxas, 2007; Knudby et al., 2013), likely due to their high conservation status in the region.

Using kernel density estimation (KDE) on research vessel trawl catch data, significant concentrations of sponges, sea pens, and large and small gorgonian corals were identified in the Scotian Shelf Biogeographic Region in 2010 (Kenchington et al., 2010). A revision of these analyses is currently underway using the most recent trawl survey data (Kenchington et al., 2016). In ecology, KDE is commonly used to identify abundance or biomass 'hotspots'. In the northwest Atlantic, this tool has been used to identify Vulnerable Marine Ecosystems (VMEs), i.e. significant concentrations of benthic structure-forming VME indicator taxa, from their broader distribution (Knudby et al., 2013; Kenchington et al., 2014). KDE is based solely on the spatial relationship between data observations and is therefore unable to extrapolate to areas where sampling has not occurred. SDMs can be a complimentary tool to KDE when evaluating potential fishing impacts or considering other management actions.

Here, we employed a specific species distribution modelling approach called random forest (RF), to predict the probability of occurrence and biomass of sponges, sea pens, and large and small gorgonian corals across DFO's Maritimes Region. Several different data sources were considered, including multispecies trawl survey catch data and *in situ* camera observations. We utilized an extensive suite of environmental predictor variables compiled specifically for the purposes of species distribution modelling in this region. With the exception of small gorgonian corals, these taxonomic groups are also considered benthic EBSA's as identified in Kenchington (2014), and all groups are considered VME indicators by the Northwest Atlantic Fisheries Organization (NAFO; NAFO, 2013). We also modelled the Russian Hat sponge *Vazella pourtalesi*, a rare VME indicator species that forms the largest known monospecific sponge ground of its kind in Emerald Basin on the Scotian Shelf. In 2013, two areas on the eastern Scotian Shelf were closed to bottom contact fishing to protect high concentrations of *V. pourtalesi* under DFO's Policy to Manage the Impacts of Fishing on Sensitive Benthic Areas. Having species distribution models for this species will give a greater picture of the extent to which these closed areas are effective. Aside from providing continuous prediction maps for the entire Maritimes Region that will be useful in ecosystem management decision-making processes, the results in this report could be used to refine the outer boundaries of the significant concentrations as identified by kernel density estimation and identify new areas that are were not sampled by the trawl surveys.

METHODOLOGY

Study Area

The Maritimes Region, one of DFO's six administrative regions across Canada, was used as the boundary for species distribution modelling in this report (Figure 1). This study area encompasses the entire Scotian Shelf and Bay of Fundy and is delimited by the Canadian Maritime Boundary to the west in Gulf of Maine, the 200 nautical mile Exclusive Economic Zone (EEZ) in the south, the Placentia Bay-Grand Bank Large Ocean Management Area in the east, and the Gulf Region MPA Network Planning Boundary in the north. A 5-km buffer was placed around all land to avoid its inclusion in the models. The total area covered in the study extent is approximately 459,139 km^2 based on a NAD 1983 UTM Zone 20N projection.

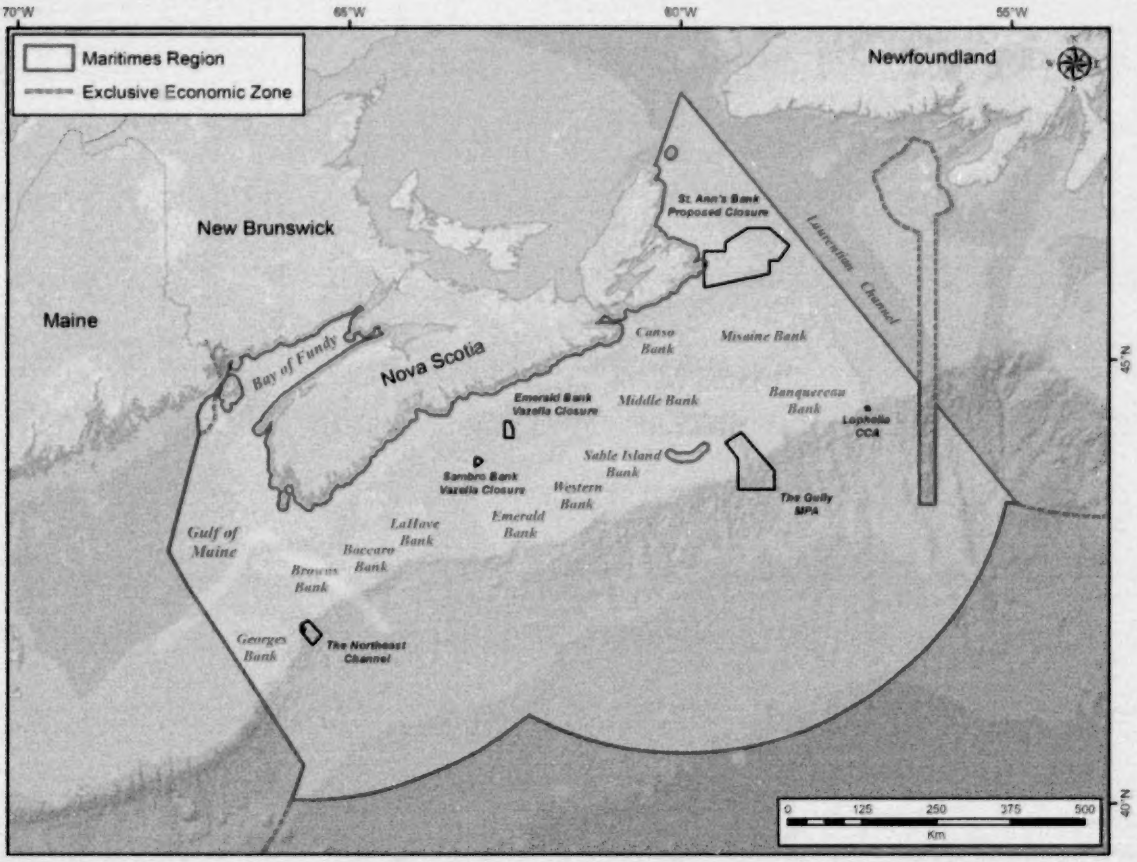

Figure 1. Extent of the DFO Maritimes Region boundary used for species distribution modelling. Place names and location of the Gully MPA, St. Ann's Bank Proposed Closure, and other areas closed to protect corals and sponges are indicated.

Environmental Data

Sixty-six environmental variables derived from various sources and native spatial resolutions were used as predictor variables in the random forest models (Table 1). Variables were chosen based on their availability and assumed relevance to the distribution of benthic fauna. Bathymetry was derived from the Canadian Hydrographic Service (CHS) Atlantic Bathymetry Compilation (ABC). This data is the highest resolution bathymetry available for the entire study area, with a horizontal resolution of up to 64 m in some areas. In the Maritimes Region the data are resolved to 15 arc-seconds, which is equivalent to approximately 500 m. Slope in degrees was derived from the depth raster using the 'Slope' tool in ArcMap's Spatial Analyst toolbox, ArcMap version 10.2.2 (ESRI, 2011). All other environmental variables were derived from long-term modelled oceanographic or remote-sensing data and were spatially interpolated across the study area using ordinary kriging in ArcMap. Specific details on data sources and methodology used for the spatial interpolation of these variables are documented in a separate technical report (Beazley et al., in prep, although see Beazley et al., 2016 for information on the same environmental data sources and variables for the Estuary and Gulf of St. Lawrence). Only variables that were spatially interpolated with reasonable confidence were used in this report, and as a result a number of available data layers (e.g. dissolved oxygen, silicate, etc.) were not considered. All predictor layers were displayed in raster format with geographic coordinates using the WGS 1984 datum and a ~0.012° cell size (approximately equal to 1 km in the Maritimes Region).

Response Data

Species composition of the five taxonomic groups modelled in this report is presented in Table 2. For each group, random forest models were generated on presence-absence records derived from catch data collected from DFO research vessel multispecies trawl surveys. Trawl surveys in the Maritimes Region were conducted on the CCGS *Alfred Needler*, *Wilfred Templeman*, or *Teleost* and followed a stratified random design (Tremblay et al., 2007). All DFO invertebrate catch data were stored in the Maritimes Region Virtual Data Centre (VDC) (http://marvdc.bio.dfo.ca/pls/vdc/mwmfdweb.auth). Data from 1999 to March 2015 were extracted from the VDC for all taxonomic groups, coinciding with the year that selected invertebrates were recorded more systematically in the surveys (Tremblay et al., 2007). Tows were conducted primarily using Western IIA trawl gear, although other gear types (i.e. Campelen and US 4 seam bridle 3 trawls) were also used in the region. Absences records were created from null (zero) catches that occurred in the same surveys.

Table 1. Summary of the 66 environmental variables used as predictor variables in random forest modelling. N/A = Not applicable.

Variable	Data source	Temporal range	Unit	Native resolution
Depth	CHS-ABC	N/A	metres	15 arc-sec.
Slope	CHS-ABC	N/A	degrees	15 arc-sec.
Bottom Salinity Mean	GLORYS2V1	1993 - 2011	N/A	¼ °
Bottom Salinity Average Minimum	GLORYS2V1	1993 - 2011	N/A	¼ °
Bottom Salinity Average Maximum	GLORYS2V1	1993 - 2011	N/A	¼ °
Bottom Salinity Average Range	GLORYS2V1	1993 - 2011	N/A	¼ °
Bottom Temperature Mean	GLORYS2V1	1993 - 2011	°C	¼ °
Bottom Temperature Average Minimum	GLORYS2V1	1993 - 2011	°C	¼ °
Bottom Temperature Average Maximum	GLORYS2V1	1993 - 2011	°C	¼ °
Bottom Temperature Average Range	GLORYS2V1	1993 - 2011	°C	¼ °
Bottom Current Speed Mean	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Bottom Current Speed Average Minimum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Bottom Current Speed Average Maximum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Bottom Current Speed Average Range	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Bottom Shear Mean	GLORYS2V1	1993 - 2011	Pa	¼ °
Bottom Shear Average Minimum	GLORYS2V1	1993 - 2011	Pa	¼ °
Bottom Shear Average Maximum	GLORYS2V1	1993 - 2011	Pa	¼ °
Bottom Shear Average Range	GLORYS2V1	1993 - 2011	Pa	¼ °
Surface Salinity Mean	GLORYS2V1	1993 - 2011	N/A	¼ °
Surface Salinity Average Minimum	GLORYS2V1	1993 - 2011	N/A	¼ °

Surface Salinity Average Maximum	GLORYS2V1	1993 - 2011	N/A	¼ °
Surface Salinity Average Range	GLORYS2V1	1993 - 2011	N/A	¼ °
Surface Temperature Mean	GLORYS2V1	1993 - 2011	°C	¼ °
Surface Temperature Average Minimum	GLORYS2V1	1993 - 2011	°C	¼ °
Surface Temperature Average Maximum	GLORYS2V1	1993 - 2011	°C	¼ °
Surface Temperature Average Range	GLORYS2V1	1993 - 2011	°C	¼ °
Surface Current Speed Mean	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Surface Current Speed Average Minimum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Surface Current Speed Average Maximum	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Surface Current Speed Average Range	GLORYS2V1	1993 - 2011	$m\ s^{-1}$	¼ °
Maximum Average Mixed Layer Depth Fall	GLORYS2V1	1993 - 2011	metres	¼ °
Maximum Average Mixed Layer Depth Winter	GLORYS2V1	1993 - 2011	metres	¼ °
Maximum Average Mixed Layer Depth Spring	GLORYS2V1	1993 - 2011	metres	¼ °
Maximum Average Mixed Layer Depth Summer	GLORYS2V1	1993 - 2011	metres	¼ °
Fall Chlorophyll *a* Mean	MODIS Level I	2002 - 2012	$mg\ m^{-3}$	2 km
Fall Chlorophyll *a* Minimum	MODIS Level I	2002 – 2012	$mg\ m^{-3}$	2 km
Fall Chlorophyll *a* Maximum	MODIS Level I	2002 – 2012	$mg\ m^{-3}$	2 km
Fall Chlorophyll *a* Range	MODIS Level I	2002 – 2012	$mg\ m^{-3}$	2 km
Spring Chlorophyll *a* Mean	MODIS Level I	2002 – 2012	$mg\ m^{-3}$	2 km
Spring Chlorophyll *a* Minimum	MODIS Level I	2002 – 2012	$mg\ m^{-3}$	2 km
Spring Chlorophyll *a* Maximum	MODIS Level I	2002 – 2012	$mg\ m^{-3}$	2 km
Spring Chlorophyll *a* Range	MODIS Level I	2002 – 2012	$mg\ m^{-3}$	2 km
Summer Chlorophyll *a* Mean	MODIS Level I	2002 – 2012	$mg\ m^{-3}$	2 km

Summer Chlorophyll *a* Minimum	MODIS Level I	2002 – 2012	mg m^{-3}	2 km
Summer Chlorophyll *a* Maximum	MODIS Level I	2002 – 2012	mg m^{-3}	2 km
Summer Chlorophyll *a* Range	MODIS Level I	2002 – 2012	mg m^{-3}	2 km
Annual Chlorophyll *a* Mean	MODIS Level I	2002 – 2012	mg m^{-3}	2 km
Annual Chlorophyll *a* Minimum	MODIS Level I	2002 – 2012	mg m^{-3}	2 km
Annual Chlorophyll *a* Maximum	MODIS Level I	2002 – 2012	mg m^{-3}	2 km
Annual Chlorophyll *a* Range	MODIS Level I	2002 – 2012	mg m^{-3}	2 km
Fall Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Fall Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Fall Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Fall Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Spring Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km

Summer Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Summer Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Mean	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Minimum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Maximum	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km
Annual Primary Production Average Range	SeaWiFS Level-3 with other input parameters	2006 – 2010	mg C m^{-2} day^{-1}	9 km

Certain areas in the Maritimes Region where hard bottom occurs are avoided during annual stock assessment surveys to prevent the loss of gear. Thus species distribution models based solely on data obtained from these surveys may be poor predictors of presence for certain benthic organisms, such as large branching corals that proliferate in areas where hard substrate occurs. For such taxonomic groups whose distribution we felt was not fully sampled by the multispecies stock assessment surveys, or when the number of trawl records for a group was insufficient for producing accurate predictions of distribution, we ran additional random forest models using trawl survey data augmented with data from other sources: 1) *in situ* benthic imagery observations from scientific surveys, 2) DFO scallop stock assessment surveys, and 3) commercial records from the Fisheries Observer Program (FOP). VDC multispecies trawl survey records using gear types other than Western IIA were also considered. Note that absence records were not generated for the benthic imagery observations or commercial trawl records. Combining data from different gear types introduces bias through differences in catchability and

Table 2. Species composition in each of the five taxonomic groups modelled using random forest. Also shown are the VDC codes used for data entry into the VDC.

Taxonomic Group	Species/Taxon	VDC Taxon Code
Sponges (Porifera)	Phylum Porifera	8600
	Geodia spp.	8364
	Polymastia sp.	8610
	Rhizaxinella sp.	8356
	Vazella pourtalesi	8601
***Vazella pourtalesi* (Russian Hat sponge)**	*Vazella pourtalesi*	8601
Sea Pens (Pennatulacea)	Order Pennatulacea	8318
	Anthoptilum grandiflorum	8361
	Funiculina quadrangularis	8359
	Halipteris sp.	8363
	Pennatula borealis	8360
Large Gorgonian Corals	*Acanthogorgia armata*	8326
	Keratoisis ornata	8325
	Paragorgia arborea	8323
	Primnoa resedaeformis	8322
Small Gorgonian Corals	*Acanella arbuscula*	8329
	Chrysogorgia agassizii	8338
	Radicipes gracilis	8330

may affect model performance. However, for presence-absence models with data matching the above conditions, we felt that the use of mixed data collection methods was justified. We did not extend this to models of biomass (see below).

Other data sources, such as local ecological knowledge records (see Gass (2002) and Breeze et al. (1997) for such reports for the region), the NOAA Deep-Sea Coral Data Portal, and those from other scientific missions were also considered for each taxonomic group and used for model validation in some cases. For sponge only, an additional 3933 observer records were obtained from Scotia-Fundy commercial surveys conducted between 1980 and 2001 using dredge, gillnets, lines, shrimp trawl, bottom otter trawl gear (S. Fuller, Ecology Action Centre, Halifax, NS; pers. comm.). Note that there may be overlap between some museum records from the NOAA Data Portal and those from the Breeze et al. (1997) report. Details on the data sources used for random forest modelling are provided separately for each taxonomic group in the Results section below.

The presence-absence records used in each random forest model (see below) were filtered so that only one presence or absence occurred within a single environmental data raster cell (~1 km). Presence records took precedence over an absence record when both occurred within the same raster cell.

Biomass (kg) data associated with the DFO multispecies trawl survey records were also extracted from the VDC. To avoid introducing any bias related to differences in catchability between gear types, only biomass data obtained from a single gear type (Western IIA trawls) were used in the random forest models. For each taxonomic group, weights were averaged across multiple tows occurring within the same environmental raster cell.

Random Forest Modelling

Random forest (Breiman, 2001), is a non-parametric machine learning technique, where multiple regression or classification trees (usually $\geq$ 500) are built using random subsets of the data (Figure 2). Each tree is fit to a bootstrap sample of the biological observations (i.e. the 'in-bag' observations), and the best split at each node is selected based on a randomly-chosen subset of predictor variables. At each node a randomly-chosen predictor value splits the response data so that maximum homogeneity is achieved (for classification) or mean response with the least error (for regression). Regression trees are used for response variables consisting of continuous data, and classification trees for factor variables. RF is a robust statistical method requiring no distributional assumptions on covariate relation to the response in comparison to other classical statistical models such as generalized linear models (GLM) or generalized additive models (GAM). It can handle a large amount of input variables effectively without variable deletion

(Chen and Ishwaran, 2012) and can also account for correlation as well as interactions among variables.

Random forest can be used to predict the probability of a species' presence (for classification) or biomass (for regression) in non-sampled areas by identifying areas with similar environmental conditions to the training data. RF models were built in the statistical computing software package R

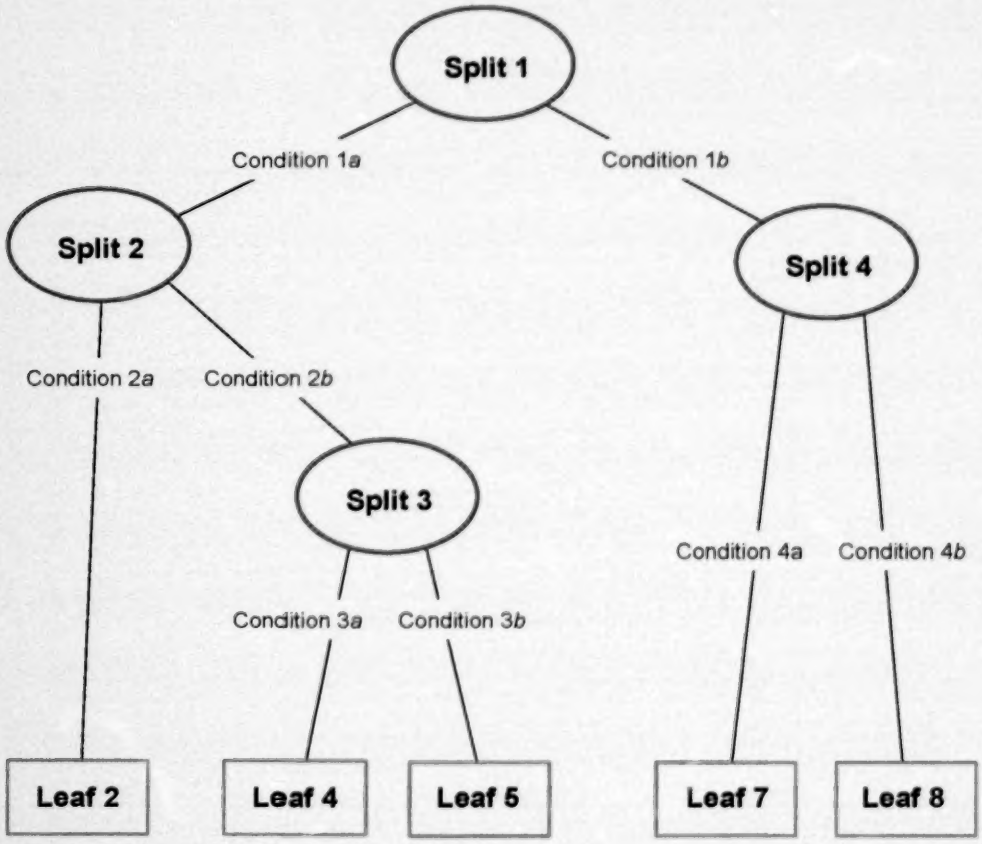

Figure 2. An example of a regression model tree (adapted from Kuhn and Johnson, 2013).

(R Core Team, 2015) using the 'randomForest' package (Liaw and Wiener, 2002). Default values were used for RF parameters, and 500 trees were constructed.

Model Evaluation

Presence-Absence Response Data – Classification Model

The catch records for some taxonomic groups are characterized by a higher number of absences relative to presences (i.e. unbalanced species prevalence, where prevalence is the proportion of presences in relation to the total dataset). The distribution of these two classes may be biased spatially and/or environmentally across the study area. Classification accuracy in random forest is prone to bias when the categorical response variable is highly imbalanced (Chen et al., 2004). This is due to over-representation of the majority class in the bootstrap sample leading to a higher frequency in which the majority class is drawn, therefore skewing predictions in that favour (Evans et al., 2011). Several different approaches have been used to address imbalanced data: 1) assign a high cost to misclassification of the minority class, 2) down-sample the majority class, and 3) up-sample the minority class (Evans et al., 2011). Although several studies suggest a balanced modelling prevalence (i.e. the proportion of presences in the dataset) of 0.5 (McPherson et al., 2004; Liu et al., 2005), this approach may result in a loss of information, particularly for rare species, and may not be necessary when the model training data is reliable and not biased spatially and/or environmentally (Jiménez-Valverde and Lobo, 2006). Another approach is to adjust the threshold used to divide the probabilistic predictions of occurrence into discrete predictions of presence or absence, to match modelling prevalence (Liu et al., 2005). The latter approach has shown to produce constant error rates and optimal model accuracy measures compared to balancing modeling prevalence (Liu et al., 2005; Hanberry and He, 2013).

Given the numerically and/or spatially biased presence and absence data of most taxonomic groups in this study, we employed two different modelling approaches and evaluated their performance. The first approach was to model the response data with a balanced species prevalence (i.e. an equal number of presences and absences) and threshold of 0.5. Here the absence records were randomly down-sampled to match the number of presences prior to modelling. In the second approach we used all presence and absence records and set the threshold equal to species prevalence. The appropriateness of each modelling approach on the response data was assessed based on the model accuracy measures (see explanation below of model accuracy measures) and the spatial pattern of the predictions of presence probability in relation to the response data.

Accuracy measures were derived from validated data using 10-fold cross validation (10 resamples over which performance estimates were obtained). In 10-fold cross validation the response data are randomly split into 10 equal-sized groups and the model is trained on a combination of 9, while validated on the remaining group. Three measures of accuracy were used to assess model performance: 1) sensitivity, 2) specificity, and 3) AUC, or Area Under the Receiver Operating Characteristic Curve. In a classification model with two classes (e.g. presence and absence), there are four possible predicted outcomes: 1) true positive, where observed presences are predicted as presences, 2) false negative, where observed presences are predicted as absences, 3) true negative, where observed absences are predicted as absences, and 4) false positive, where observed absences are predicted as presences (Fawcett, 2006).

Sensitivity measures the proportion of observed presences correctly predicted as presence (i.e. the true positive rate) (McPherson et al., 2004; Fawcett, 2006). Low sensitivity indicates high omission error (i.e. false negative rate). Specificity measures the proportion of observed absences correctly predicted as absence (i.e. the true negative rate). Low specificity indicates high commission error (i.e. the false positive rate). Both sensitivity and specificity are derived from a two-by-two confusion matrix of the tabulated predicted outcomes.

The AUC is a threshold-independent measure of model accuracy that is calculated from the combination of true positive rate (sensitivity) and false positive rate (1 – specificity), and equals the probability that the model will rank a randomly-chosen presence instance higher than a randomly-chosen absence instance (Fawcett, 2006). Its value ranges from 0 to 1, with values larger than 0.5 indicating performance better than random (Fawcett, 2006).

For models generated using a balanced species prevalence and threshold of 0.5, 10 data subsets were created with the same number of presence and absences (balanced data) and the AUC was determined by averaging the AUC values between folds within each run. The model with the highest average AUC was considered the most accurate in predicting the validated data and was used as the final model in which predicted presence probabilities of the response data were generated. The predicted outcomes from the two-by-two confusion matrices were summed across all 10 folds to give a complete confusion matrix for each model from which sensitivity and specificity were calculated. For models generated using all presence and absence data and a threshold equal to species prevalence, only one model was considered and the AUC was determined by averaging AUC values between folds. The predicted outcomes from the two-by-two confusion matrices were summed across all 10 folds to give one confusion matrix from which sensitivity and specificity were calculated.

Biomass Response Data – Regression Model

Models were validated using 10-fold cross validation. Data were split using the createFolds function in R. This function performs stratified partitioning into k groups (=folds) to better evenly distribute the biomass values across splits. Models were built using each calibrated and validated dataset and accuracy measures were calculated for each corresponding dataset. The accuracy measures used to validate the models included the goodness-of-fit statistic R^2, the Root-Mean-Square Error (RMSE) and the percentage of variance explained. RMSE was normalized to a percentage of the range of observed biomass values ($y_{max} - y_{min}$) for each specific response (NRMSE) to facilitate the comparison between responses in the different models. Cross validation gives an average of the accuracy measures used, but can also be used to estimate the variability around the mean to evaluate the stability of the model fit, and to check for the arbitrary effects from subsampling data.

Model Extrapolation

The spatial extent of the Maritimes Region reaches far beyond the Scotian Shelf and Slope, down to ~5100 m depth. Our data observations are limited to depths above ~2900 m (multispecies trawl observations are limited to depths of 1850 m and shallower). Extrapolation of model predictions to areas outside of the range of data observations may produce unreliable predictions in those areas (Elith et al., 2010). Random forest models average the decision across regression trees to predict piecewise constant functions, giving a constant value for inputs falling under each leaf. When extrapolating outside the domain of the training data, where different physical conditions from those used to train the model likely exist, random forest models predict the same value as they would for the closest value in the tree for which they had training data (Breiman et al., 1984). For each random forest model, we highlight those areas within the study extent where model predictions are extrapolated. We define areas of extrapolation as those areas where at least one environmental variable has values above or below its sampled range.

Ecological Interpretation

Ecological interpretation of the models was aided by predictor variable importance measures and partial dependence plots. In classification models, variable importance is measured as the mean decrease in Gini value, otherwise known as Gini impurity. When the response data are split into two child nodes based on a randomly-chosen variable, the data in the two descendent nodes are more homogeneous than that of the parent node. This difference in homogeneity between parent and child nodes is measured by the Gini index, where the increase in homogeneity equals a decrease in Gini value. The sum of all decreases in Gini index for each variable in each tree is averaged across all trees in the model 'forest' and then across all 10 repetitions of each model fold. The variable with the highest mean decrease in Gini value is considered the most important variable in the model. Variable importance in regression random forest is measured by the mean decrease in the residual sum of squares when the variable is included in a tree split.

Partial dependence plots using the partialPlot function in R were generated for the six most important variables. Partial dependence plots show the relationship between a particular predictor variable and the log-transformed predicted probabilities of presence for classification models or the biomass regression function for regression models. The other predictor variables are held constant at their mean observed value. Partial dependence plots are useful in showing general trends in model accuracy's dependence on the predictors (Herrick et al., 2013). For classification models, the y axis ranges from $-\infty$ to ∞ and quantifies the log-odds of a positive classification for the total range of values in x. Log-odds are logarithmic transformations of the probabilities for values in x (Hastie et al., 2005). These values were transformed to the original presence probability scale using:

$$p = \exp(y) / (1 + \exp(y)),$$

where p = the probability of presence, and y is the log-odds of presence, the standard output from the partialPlot function.

Alternative Prediction Models

Generalized additive models (GAMs) were developed to predict the biomass distribution of each taxonomic group. GAMs were developed to compare a regression approach to the machine learning random forest results and to determine whether predictions could be improved for the areas considered as extrapolated by Random forest models. Methodology and results for the GAM models are presented in Appendix 1.

RESULTS

Sponges (Porifera)

Data Sources and Distribution

Figure 3 shows the distribution of available sponge records in the Maritimes Region. Aside from scientific survey records, DFO multispecies trawl surveys using Western IIA gear (white circles) accounted for the majority of sponge records in the region. These trawl records were distributed relatively evenly across the Scotian Shelf and Slope down to 1850 m depth, but were less numerous in the Bay of Fundy region. As a result, the DFO multispecies trawl survey Western IIA data were augmented with catch records from DFO scallop stock assessment

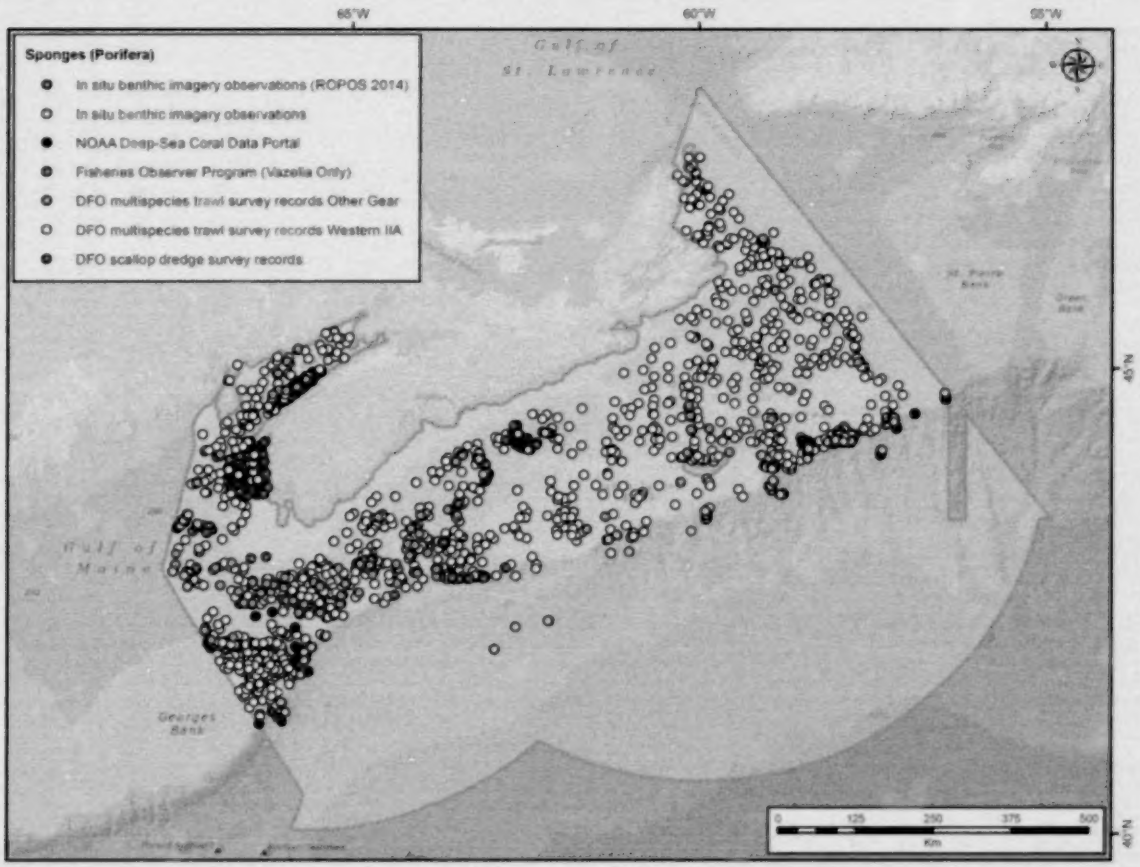

Figure 3. Available sponge presence data in the Maritimes Region from scientific survey missions, the NOAA Deep-Sea Coral Data Portal, DFO research vessel surveys, and commercial catches from the Fisheries Observer Program.

surveys conducted off Digby (Scallop Production Area (SPA) 4) and the Brier Island, Lurcher Shoal, and St. Mary's Bay area (SPA 3). These surveys used modified Digby scallop drags (Kenchington et al., 2007). The DFO multispecies survey data consists of 1174 presence and 1846 absence records collected in various years between 2001 and 2015 (Table 3; Figure 4). The scallop catch data consists of 136 presence and 6 absence records from 1997, and 107 presence and 2 absence records from 2007. The combined dataset used for modelling consists of 1417 presence and 1854 absence records (Figure 4). The highest mean biomass value from the Western IIA records (85.54 kg) occurred in Emerald Basin (Figure 5), an area dominated by the Russian Hat sponge *Vazella pourtalesi*. Several large biomass records also occurred off southwestern Nova Scotia. The area inshore of Browns Bank (Figure 1) off southwestern Nova Scotia is not surveyed (Figures 4, 5).

Table 3. Number of presence and absence records of sponge catch recorded from DFO multispecies trawl and scallop stock assessment surveys conducted between 1997 and 2015 in the Maritimes Region.

Year	Total number of presences	Total number of absences
1997	136	6
2001	1	93
2002	57	151
2005	21	167
2006	6	94
2007	213	200
2008	109	176
2009	130	136
2010	132	204
2011	152	160
2012	125	141
2013	163	189
2014	148	128
2015	24	9

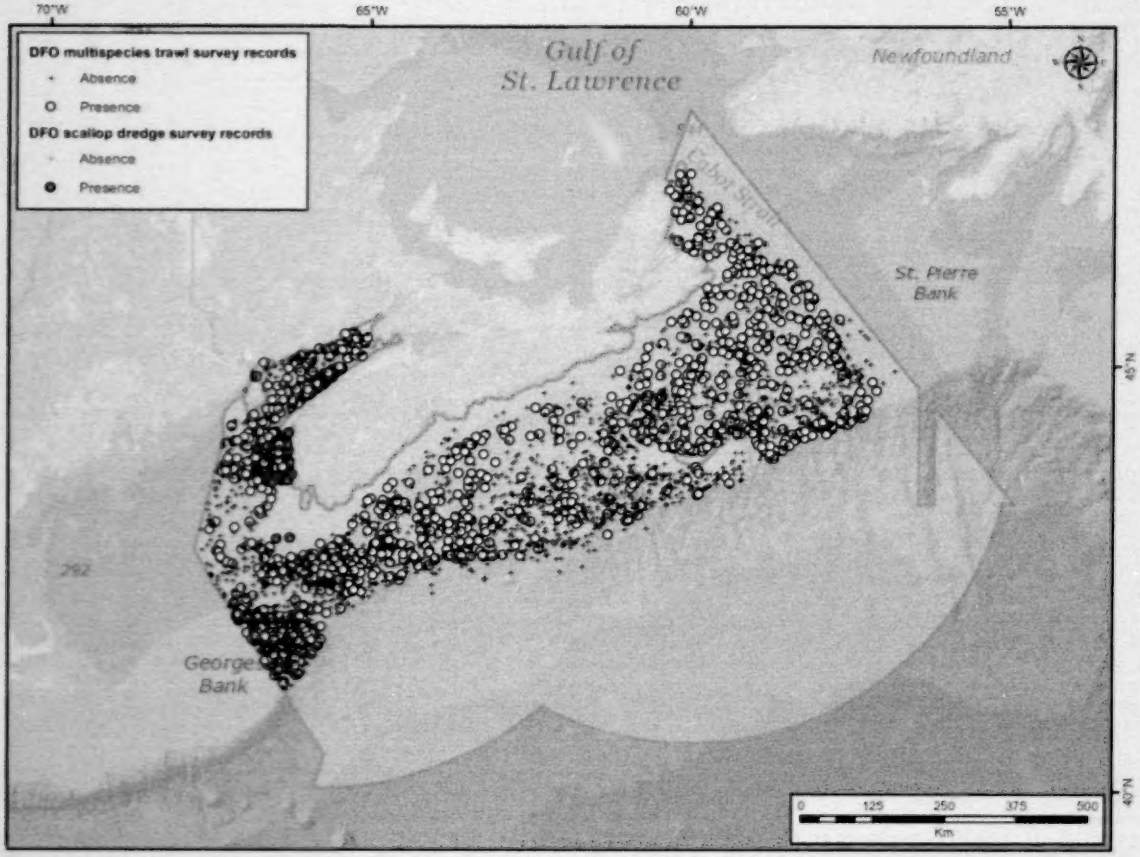

Figure 4. Presence and absence records of sponge catch recorded from DFO multispecies trawl and scallop stock assessment surveys from 1997 to 2015 within the Maritimes Region.

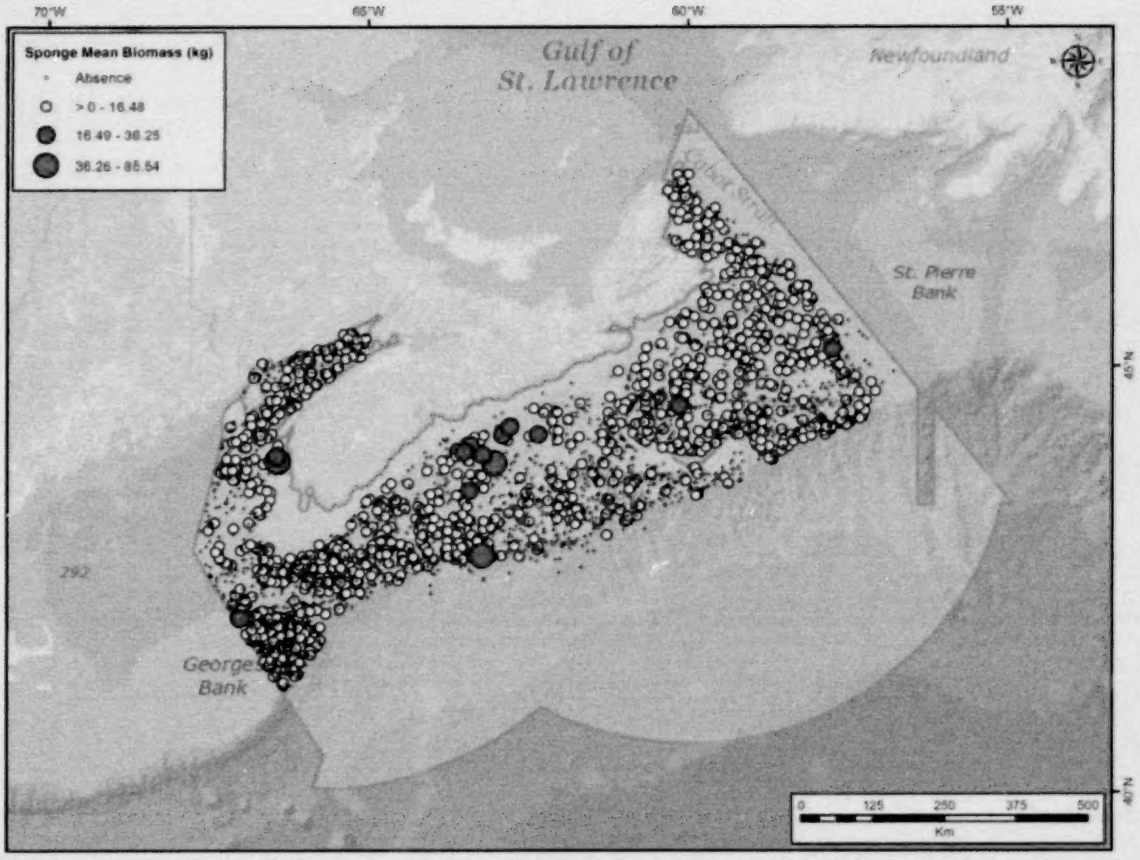

Figure 5. Mean biomass (kg) per grid cell of sponge catch recorded from DFO multispecies trawl surveys from 2001 to 2015 within the Maritimes Region. Also shown are absence records from both DFO multispecies trawl and scallop stock assessment surveys collected between 1997 and 2015.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity, and specificity) for the random forest model using a balanced species prevalence (1417 presences and 1417 absences; Model 1) are presented in Table 4. The highest mean AUC of 0.766 was associated with Model 1 and is therefore considered the optimal model for the prediction of the sponge response data. The sensitivity and specificity measures were 0.689 and 0.708, respectively. The confusion matrix of the optimal model is also presented in Table 2. Class error for both the presence and absence classes was somewhat moderate (0.311 and 0.292, respectively).

Table 4. Accuracy measures for all 10 model repetitions of 10-fold across validation of a random forest model of sponge presence-absence data collected within the Maritimes Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 1) which is considered the optimal model for predicting the presence probability of sponge in the region.

Model Run	AUC	Sensitivity	Specificity
1	**0.766**	**0.689**	**0.708**
2	0.757	0.691	0.697
3	0.760	0.688	0.707
4	0.759	0.694	0.705
5	0.761	0.699	0.701
6	0.763	0.693	0.706
7	0.749	0.688	0.694
8	0.764	0.696	0.716
9	0.756	0.697	0.691
10	0.762	0.680	0.698
Mean	**0.760**	**0.691**	**0.702**
SD	**0.005**	**0.005**	**0.007**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	**Presence**		
Absence	1003	414	1417	0.292
Presence	441	976	1417	0.311

The presence probability prediction surface of sponges is presented in Figure 6. Pockets of high presence probability were distributed across the study area, but several areas had notably high presence probability: Smokey and St. Ann's Banks off northeastern Nova Scotia (Cape Breton), Misaine Bank, and the Bay of Fundy off Digby and Brier Island. The latter two areas corresponded to the location of the additional sponge records from the DFO scallop stock assessment surveys in SPA 3 and 4 (Figure 7). Other areas of high presence probability corresponded well with the occurrence of presence points at those locations. Interestingly, the unsampled area southwest of Nova Scotia has a moderate to high presence probability of sponges.

Figure 8 shows the actual data observations (1417 presences and 1417 absences) used in Model 1. There appears to be little to no spatial bias in the presence and absences records used in the model. Areas of extrapolation are also shown in Figure 8. Small pockets of extrapolated area are distributed across the Scotian Shelf, with larger areas occurring off southwestern Nova Scotia and off northeast Cape Breton. All deep water beyond the Scotian Shelf is considered extrapolated area. The extrapolated area off southwestern Nova Scotia is predicted to have a high presence probability of sponges.

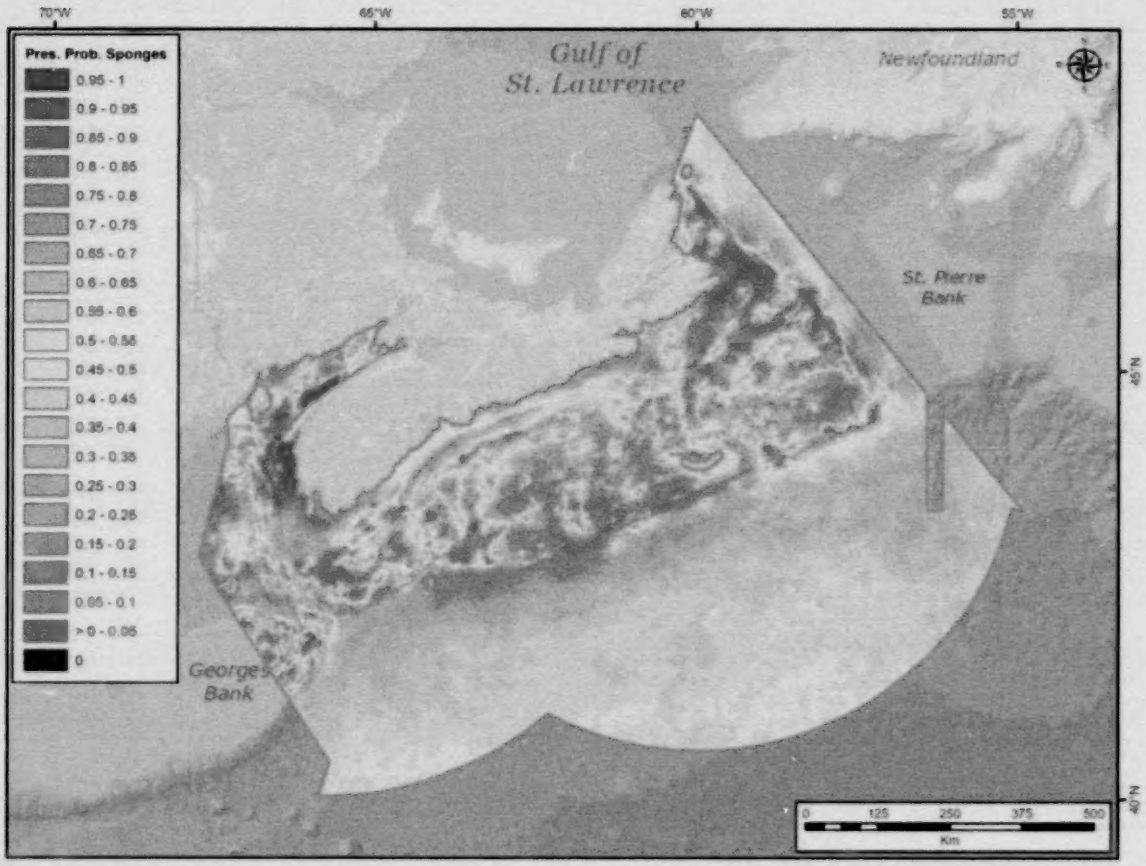

Figure 6. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of sponge presence and absence data collected from DFO multispecies trawl and scallop stock assessment surveys in the Maritimes Region between 1997 and 2015.

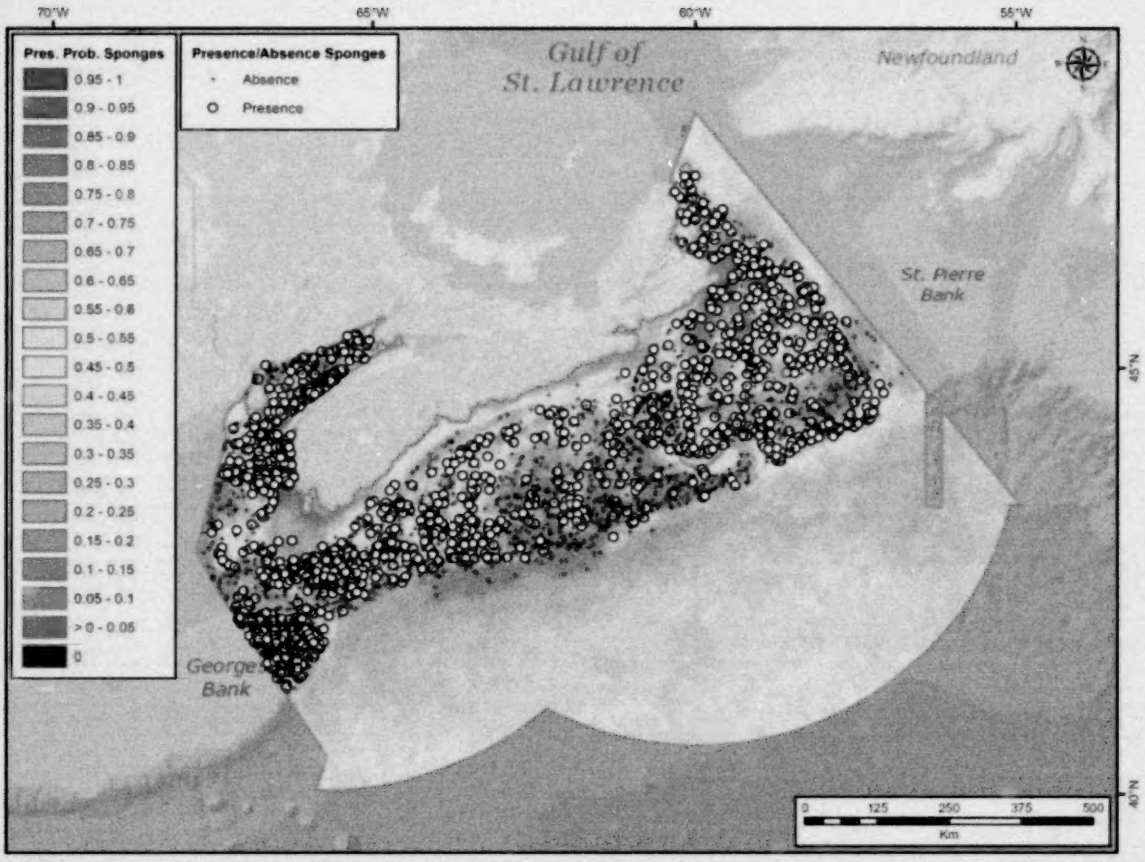

Figure 7. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of sponge presence and absence data recorded from DFO multispecies trawl and scallop stock assessment surveys in the Maritimes Region between 1997 and 2015.

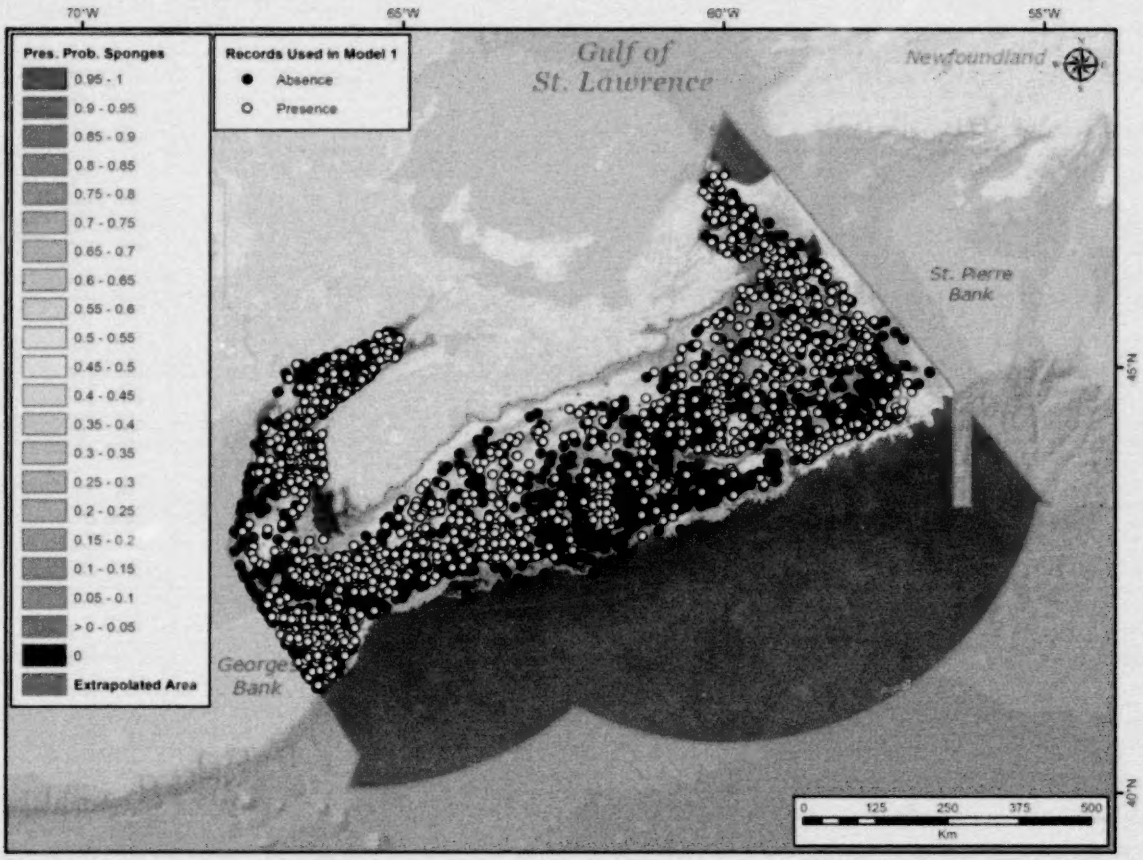

Figure 8. Map of the 2834 data observations (1417 presences and 1417 absences) of sponges used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of sponges and the areas of model extrapolation.

Of all 66 environmental predictor variables used in the model, Maximum Average Summer Mixed Layer Depth was the most important for the classification of the sponge presence and absence data (Figure 9). Prior to spatial interpolation, this variable displayed a right-skewed distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed a strong spatial pattern in data points over- and under-predicted by a normal distribution, with over-predicted points located along the coast of Nova Scotia and in the deepest regions of the study extent, and under-predicted points located across the centre of the study extent. Maximum Average Summer Mixed Layer Depth was followed more distantly in terms of its Mean Decrease in Gini Value by Depth, Surface Temperature Average Maximum, and Spring Chlorophyll *a* Mean. Chlorophyll *a* variables ranked high in the model. Partial dependence plots for the top 6 predictor variables are shown in Figure 10. The highest predicted sponge presence probabilities were associated with Maximum Average Summer Mixed Layer Depth values between 11 and 13

m. Values in this range coincided with those data points along the coast that were over-predicted by a normal distribution. However, the fit between predicted and observed values for this variable was very good, with little deviation in data points from the 1:1 reference line. Any data points over-predicted by the kriging model were within 1 m of the true values, and were still within the range of high presence probability identified in the partial plot (Figure 10). Along the Depth gradient, presence probability was highest at the shallowest depths.

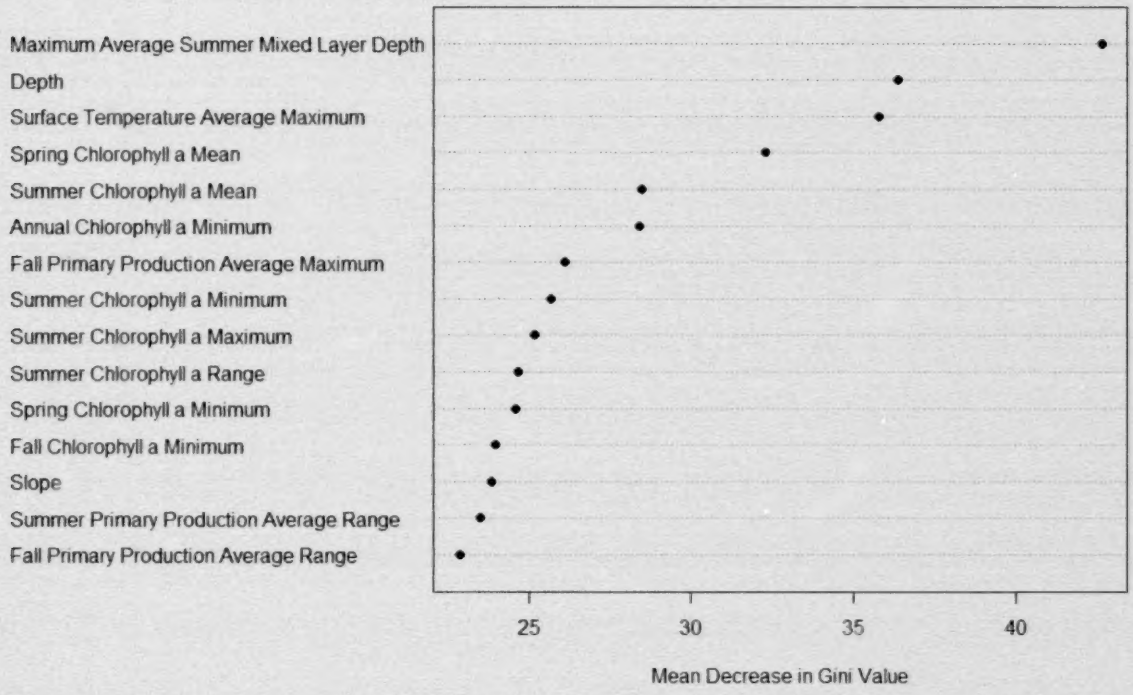

Figure 9. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting sponge presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

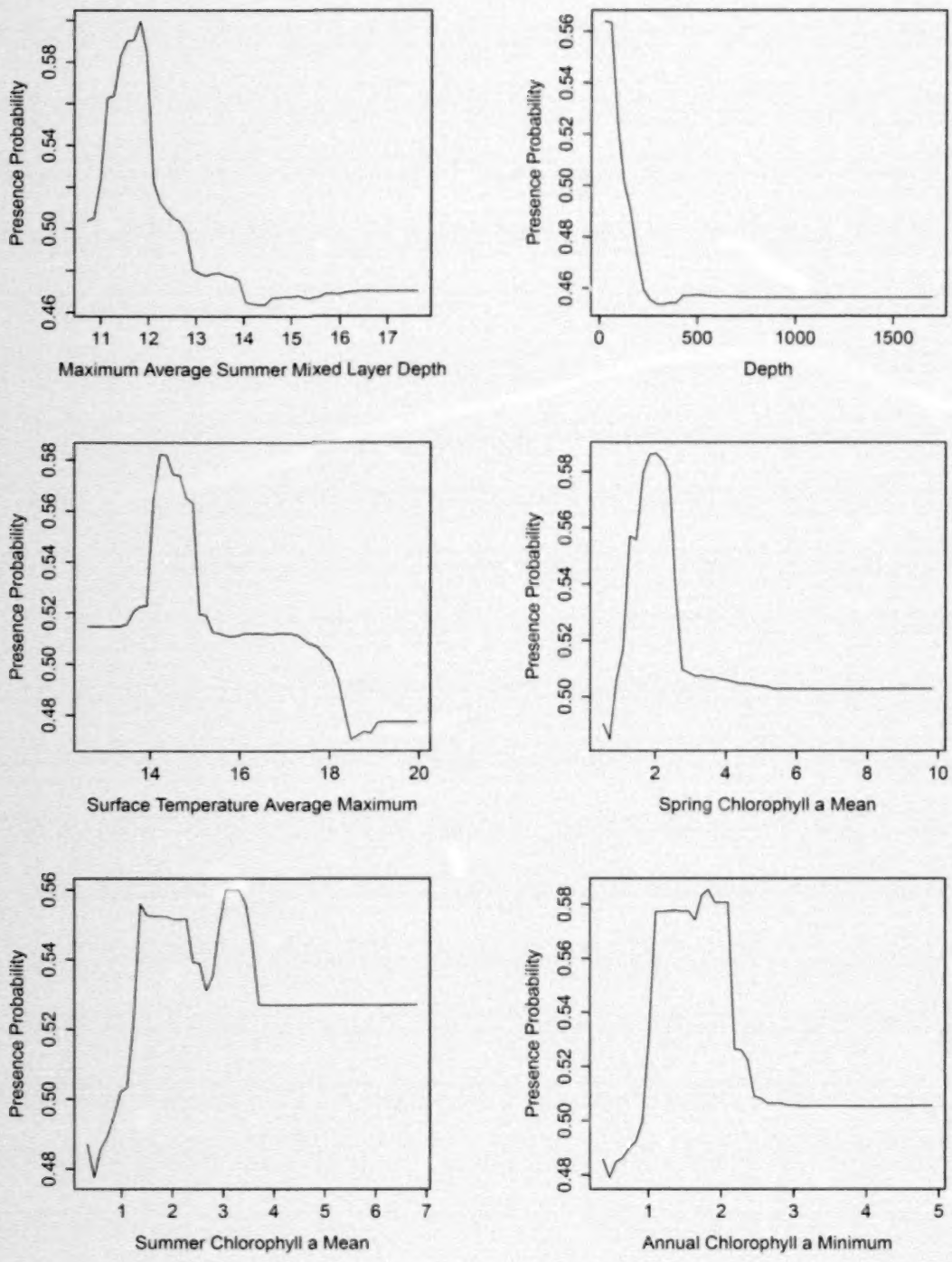
Presence Probability
0.46 0.50 0.54 0.58
11 12 13 14 15 16 17
Maximum Average Summer Mixed Layer Depth
Presence Probability
0.46 0.48 0.50 0.52 0.54 0.56
0 500 1000 1500
Depth
Presence Probability
0.48 0.50 0.52 0.54 0.56 0.58
14 16 18 20
Surface Temperature Average Maximum
Presence Probability
0.50 0.52 0.54 0.56 0.58
2 4 6 8 10
Spring Chlorophyll a Mean
Presence Probability
0.48 0.50 0.52 0.54 0.56
1 2 3 4 5 6 7
Summer Chlorophyll a Mean
Presence Probability
0.48 0.50 0.52 0.54 0.56 0.58
1 2 3 4 5
Annual Chlorophyll a Minimum

Figure 10. Partial dependence plots of the top six predictors from the optimal random forest model of sponge presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 5 shows the accuracy measures for the random forest model using all sponge presence and absence data (1417 presences and 1854 absences; Model 2) and a threshold equal to species prevalence (0.43). The average AUC calculated from Model 2 was 0.766, higher than that of Model 1. Class error of the absence class was lower than Model 1, while class error for the presence class was the same between the two models.

The surface of predicted presence probability of sponges generated from Model 2 is nearly identical to that of Model 1 (Figure 11). The area of high presence probability on St. Ann's Bank is slightly reduced in this model. The model does not appear to predict areas of presence far beyond the location of presence records (Figure 12), likely due to the inclusion of all absence records in the model. Areas of extrapolation are shown in Figure 13, likely due to the even spatial distribution of the presence-absence data. The locations of extrapolated area in this model were nearly identical to that of Model 1. Figure 14 depicts the classification of sponge presence probability into presence and absence categories based on the prevalence threshold of 0.43. In this map, all presence probability values generated from Model 2 that were greater than 0.43 were classified as presence, while values less than 0.43 were classed as absence. The largest areas classified as sponge presence were located on St. Ann's Bank and off southwestern Nova Scotia. Much of central and outer Scotian Shelf was classified as absence of sponges.

Table 5. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of sponges within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.765		**Absence**	**Presence**				
2	0.757	**Absence**	1317	537	1854	0.290	0.709	0.710
3	0.774	**Presence**	413	1004	1417	0.292		
4	0.802							
5	0.756							
6	0.779							
7	0.769							
8	0.745							
9	0.766							
10	0.745							
Mean	**0.766**							
SD	**0.017**							

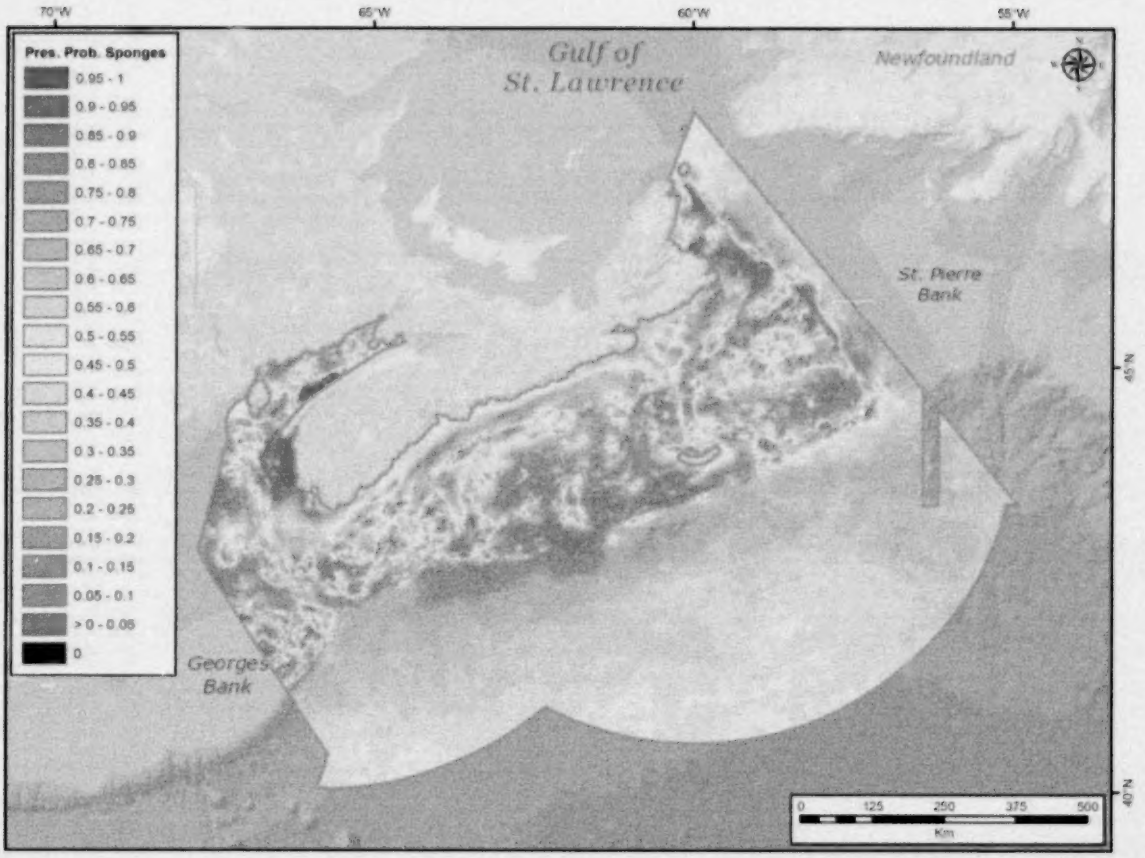

Figure 11. Predictions of presence probability (Pres. Prob.) of sponges based on a random forest model on unbalanced presence and absence sponge catch data collected from DFO multispecies trawl and scallop stock assessment surveys conducted in the Maritimes Region between 1997 and 2015.

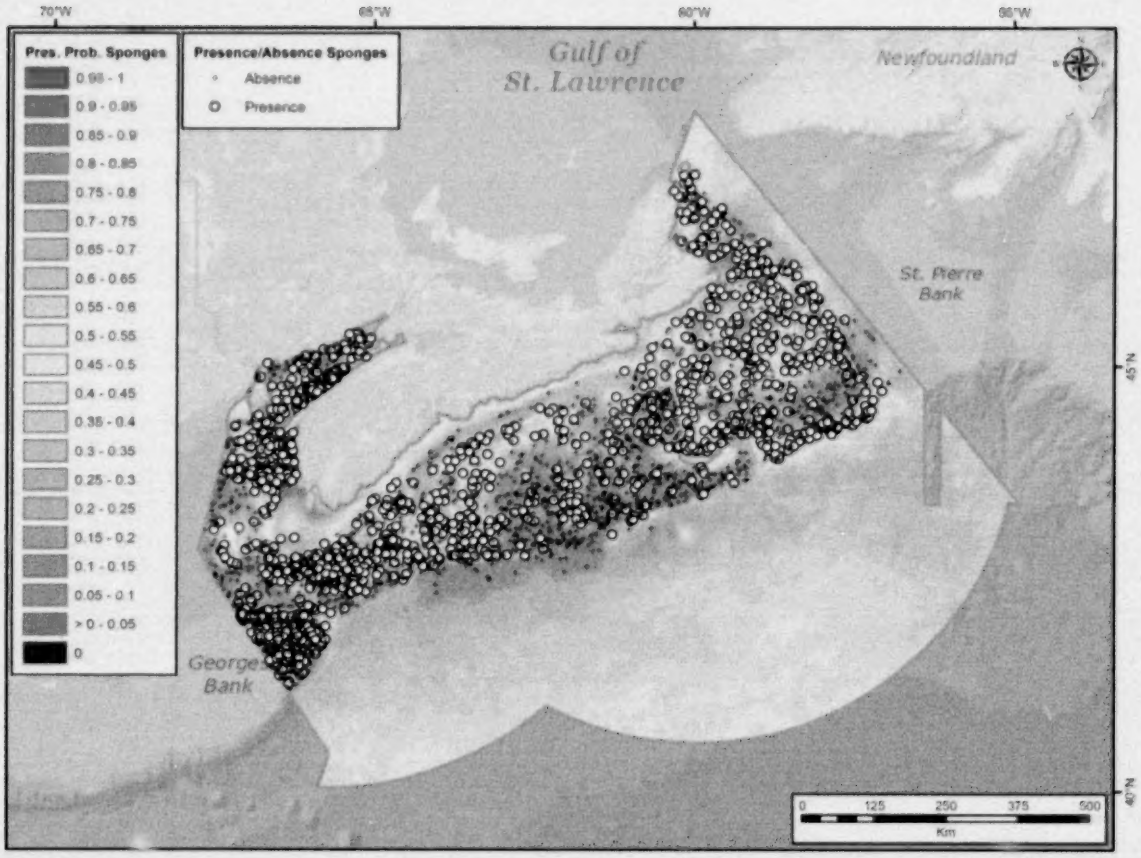

Figure 12. Presence and absence observations and predictions of presence probability (Pres. Prob.) of sponges based on a random forest model on unbalanced presence and absence sponge catch data collected from DFO multispecies trawl and scallop stock assessment surveys conducted in the Maritimes Region between 1997 and 2015.

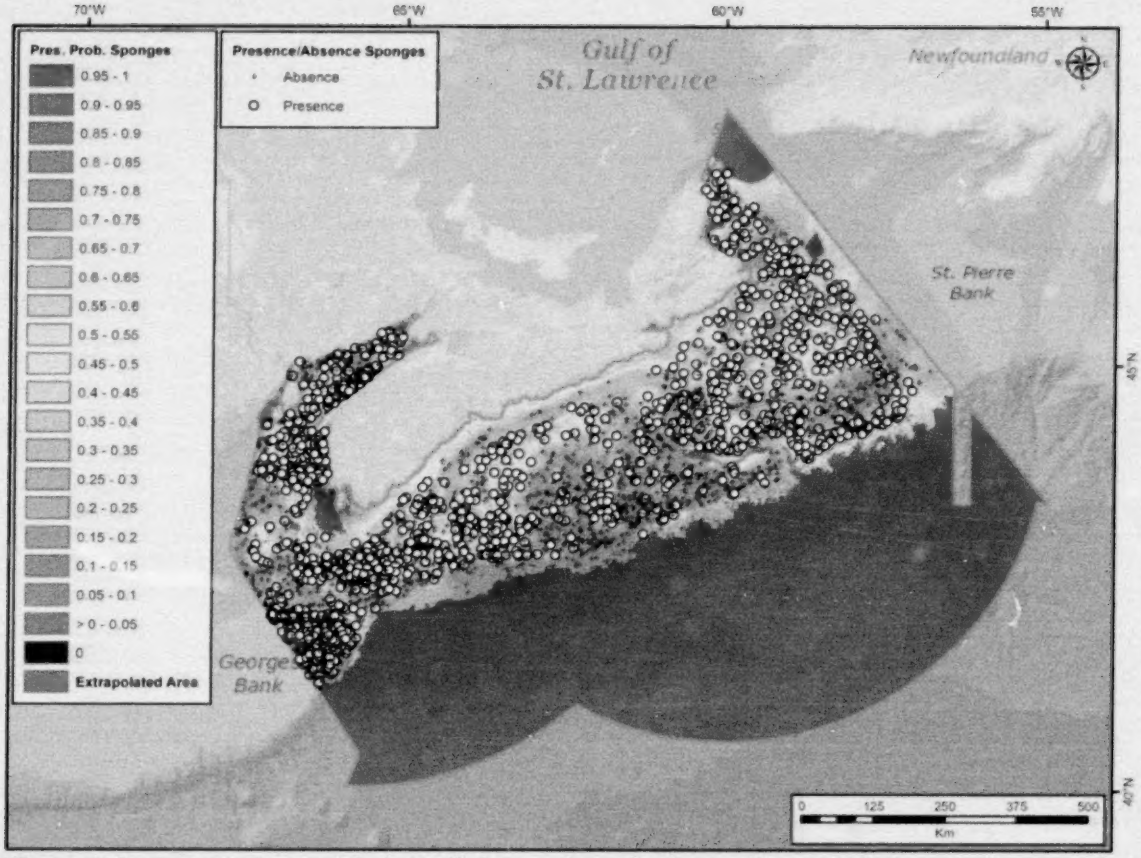

Figure 13. Areas of extrapolation of the random forest model on unbalanced presence and absence sponge catch data collected within the Maritimes Region between 1997 and 2015. Also shown are the sponge presence and absence observations and predictions of presence probability (Pres. Prob.).

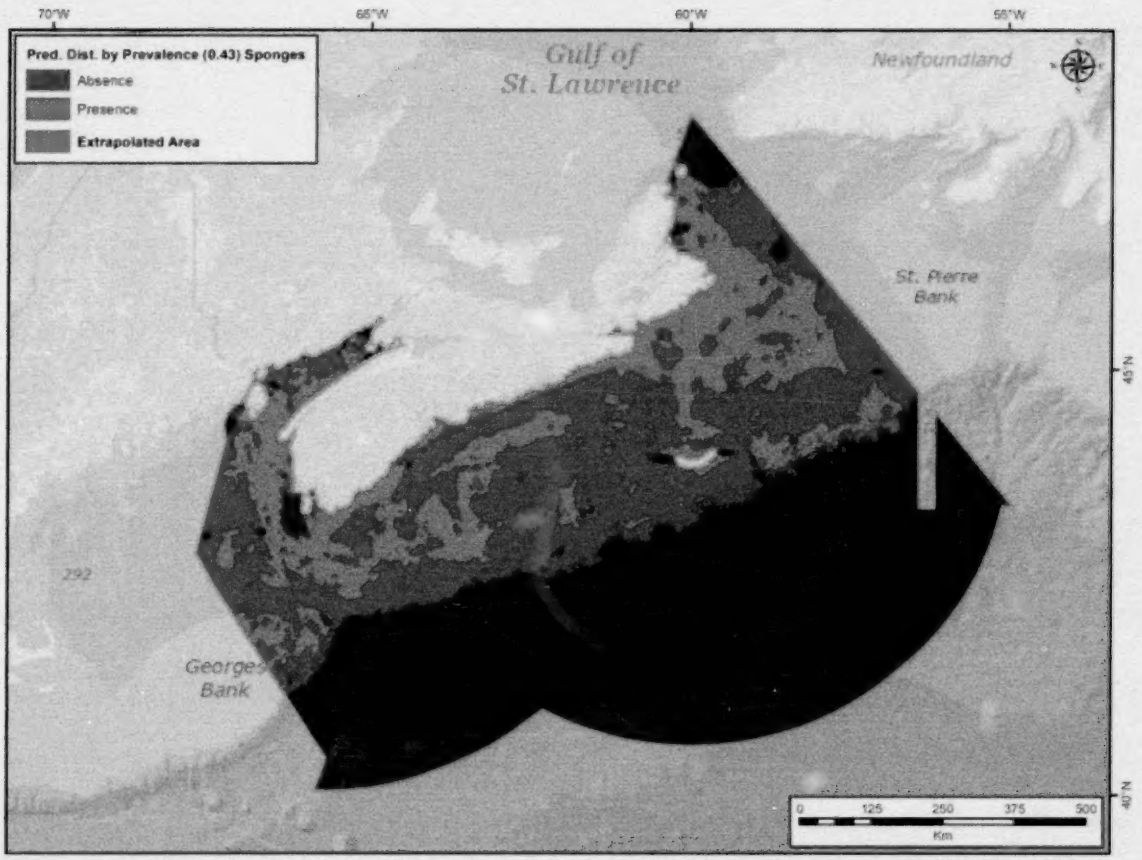

Figure 14. Predicted distribution (Pred. Dist.) of sponges in the Maritimes Region based on the prevalence threshold of 0.43 of sponge presence and absence data used in Model 2. Also shown are the areas of model extrapolation (grey polygon may appear red or blue).

The order of importance of the top environmental predictor variables in Model 2 (Figure 15) was nearly identical to that of Model 1, with Maximum Average Summer Mixed Layer Depth, Depth, and Surface Temperature Average Maximum holding the top three spots. The order of the remaining variables changed slightly from Model 1. The partial dependence of sponge presence and absence data on the top 6 predictor variables is shown in Figure 16. Sponge presence probability was highest at Maximum Average Summer Mixed Layer Depth values between ~11 and 12 m. Similar to Model 1, presence probability was highest at the shallowest depths along the Depth gradient.

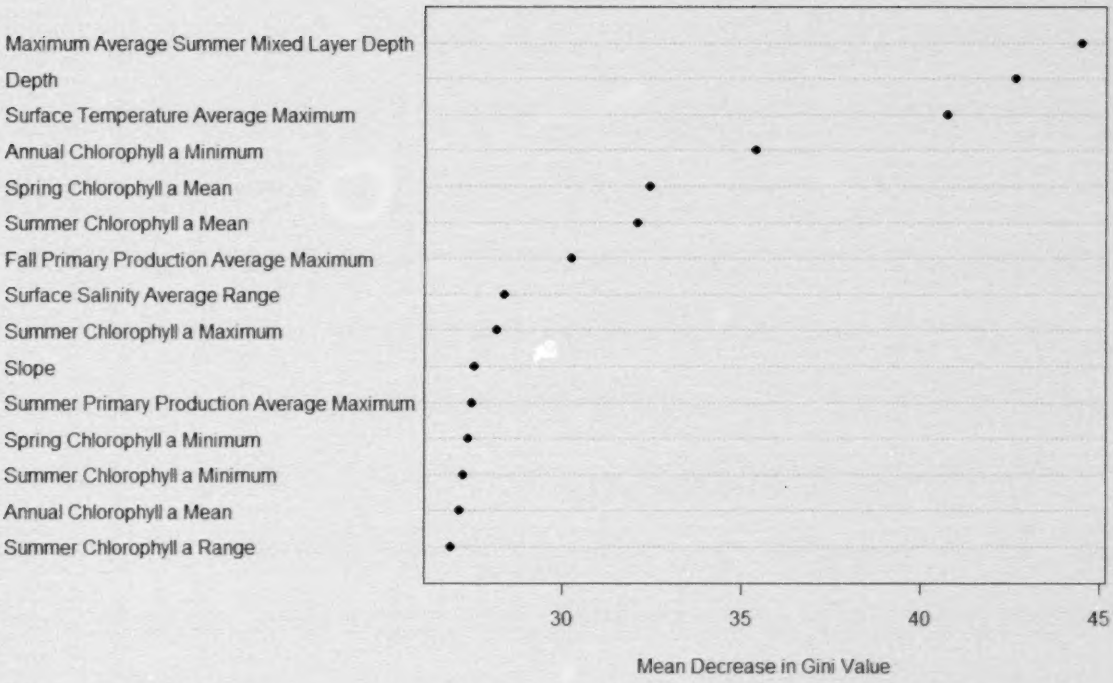

Figure 15. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced sponge presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

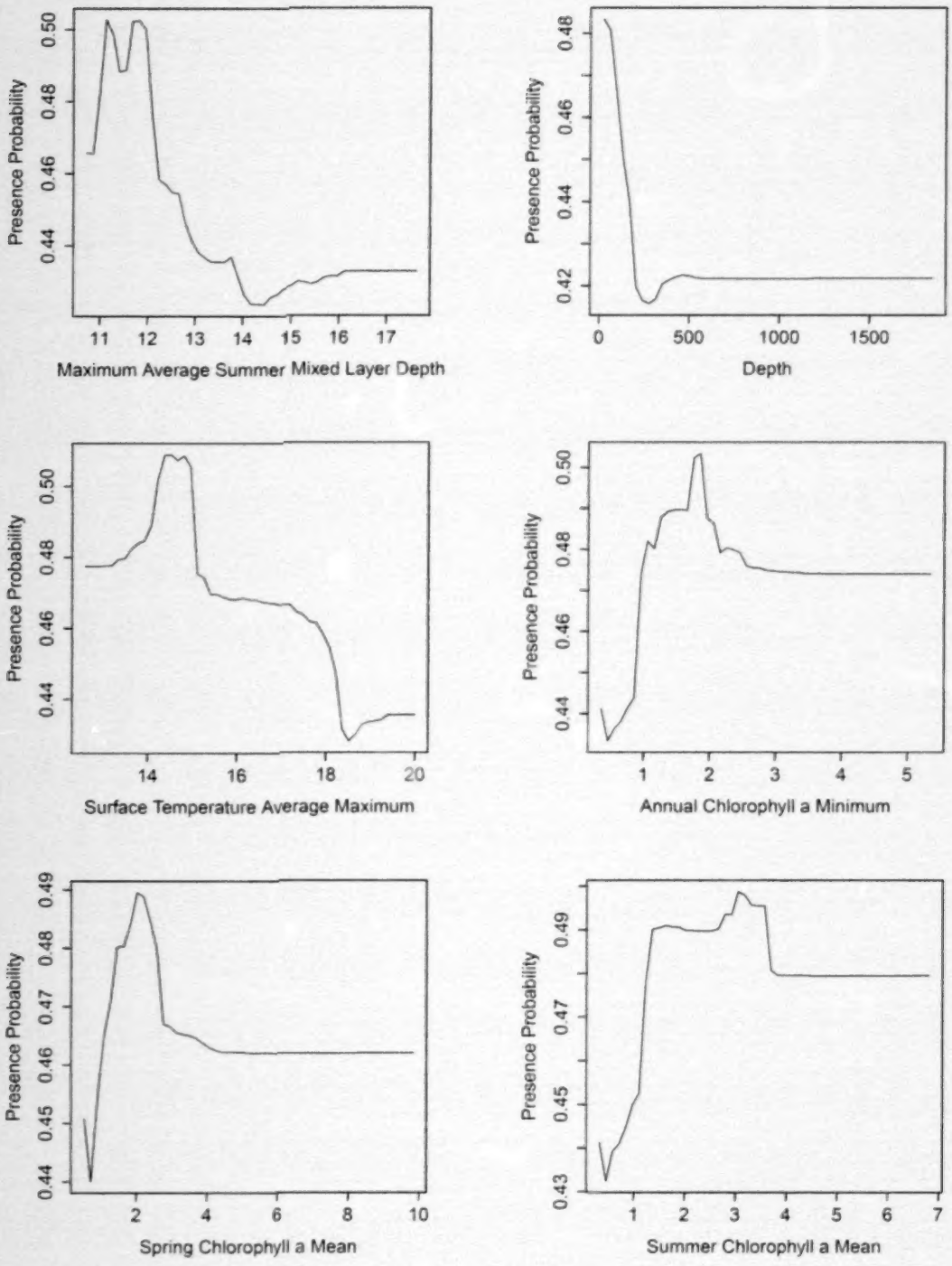

Figure 16. Partial dependence plots of the top six predictors from the random forest model of sponge unbalanced presence and absence data collected within the Maritimes Region study, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using a balanced species prevalence and threshold equal to 0.5 (Model 1) was chosen as the best predictor of sponge distribution in the Maritimes Region. Although model accuracy measures were slightly better for Model 2, the presence probability surfaces were nearly identical between both models. The selection of Model 1 allows for the use of presence probability over presence or absence classification based on the prevalence threshold.

Validation of Selected Model Using Independent Data

Figure 17 shows the predicted presence probability of sponges generated from the selected model (Model 1) at the location of additional sponge records not used in the model from *in situ* benthic imagery observations and DFO multispecies trawl surveys using Campelen and US 4 seam 3 bridle trawl gear. There is relatively good congruence between the location of sponge records from the *in situ* benthic imagery observations and areas of high presence predicted by the model. Many of the camera observations were concentrated in the Northeast Channel and along the eastern slope and its canyons where there is a relatively high predicted presence probability of sponges (top map of Figure 17). No science survey records exist off Cape Breton or southwestern Nova Scotia where the highest predicted presence probabilities occurred. Several records occurred in deeper waters off the shelf in an area considered extrapolated by the model.

The sponge records from DFO multispecies trawl surveys using Campelen and US 4 seam 3 bridle trawl gear were distributed mainly in the western portion of the study area on Browns, Baccaro, and LaHave Banks, and in Bay of Fundy where there was a relatively high presence probability of sponges predicted by the model. There was also good congruence between the location of observer records from the Scotia Fundy survey and areas of high presence predicted by the model. A notable exception is a cluster of records on Banquereau Bank that were predicted with low probability.

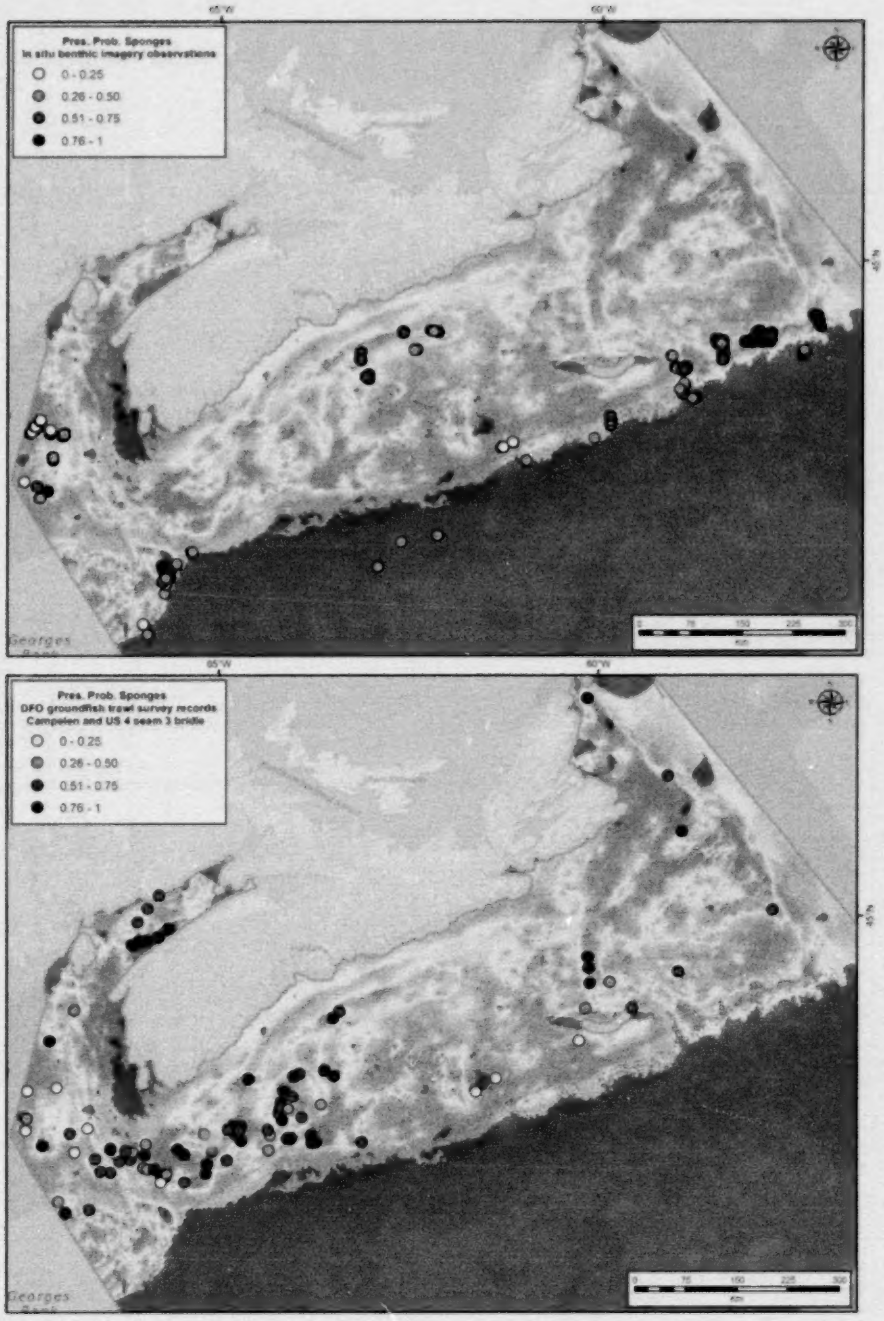

Figure 17. Validation of sponge presence probability from Model 1 using independent data. Presence probability values were extracted to the location of sponge records from DFO multispecies trawl surveys using Campelen and US 4 seam 3 bridle trawl gear (top map), *in situ* benthic imagery observations from scientific surveys (bottom map), and Scotia Fundy observer records (see next page).

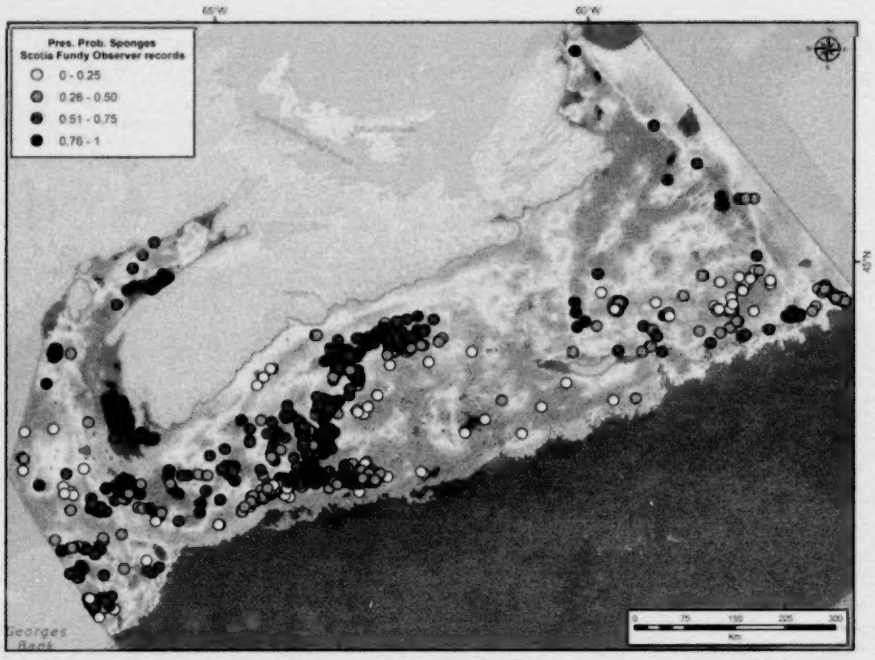

Figure 17. continued.

Prediction of Sponge Biomass Using Random Forest

The accuracy measures of the regression random forest model of mean sponge biomass per grid cell from DFO multispecies trawl surveys are presented in Table 6. The highest R^2 was 0.4588 with a mean of 0.130 ± 0.138 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.030 ± 0.013 SD. The highest percentage variance explained was 8.51%, however, half of the model folds had a negative variance explained indicating poor predictive performance of the model.

Figures 18 and 19 show the surface of sponge biomass (kg) predictions per grid cell generated from the random forest model. The majority of the spatial extent was predicted to have low sponge biomass. However, the few areas of high biomass predicted by the model coincide well with the locations of high biomass records in Emerald Basin (Figure 19).

Table 6. Accuracy measures from 10-fold cross validation of a random forest model of average sponge biomass (kg) per grid cell recorded from DFO multispecies trawl surveys in the Maritimes Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.114	1.834	0.021	-0.73
2	0.028	1.543	0.018	1.84
3	0.459	3.317	0.039	-3.03
4	0.032	1.990	0.023	0.63
5	0.016	4.919	0.058	8.51
6	0.110	2.757	0.032	-0.76
7	0.275	2.258	0.026	-4.64
8	0.052	3.397	0.040	1.01
9	0.142	2.876	0.034	0.14
10	0.071	1.083	0.013	-0.85
Mean	**0.130**	**2.597**	**0.030**	**0.21**
SD	**0.138**	**1.112**	**0.013**	**3.49**

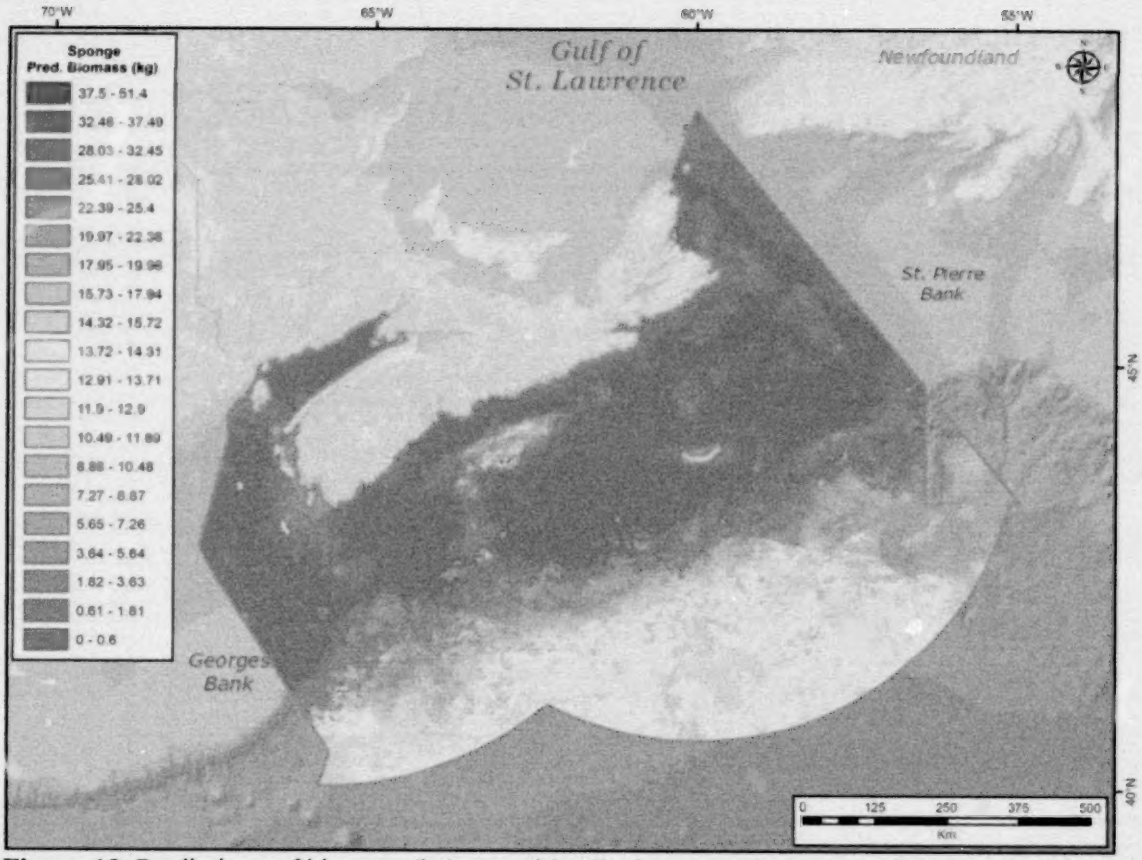

Figure 18. Predictions of biomass (kg) per grid cell of sponges from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2001 and 2015.

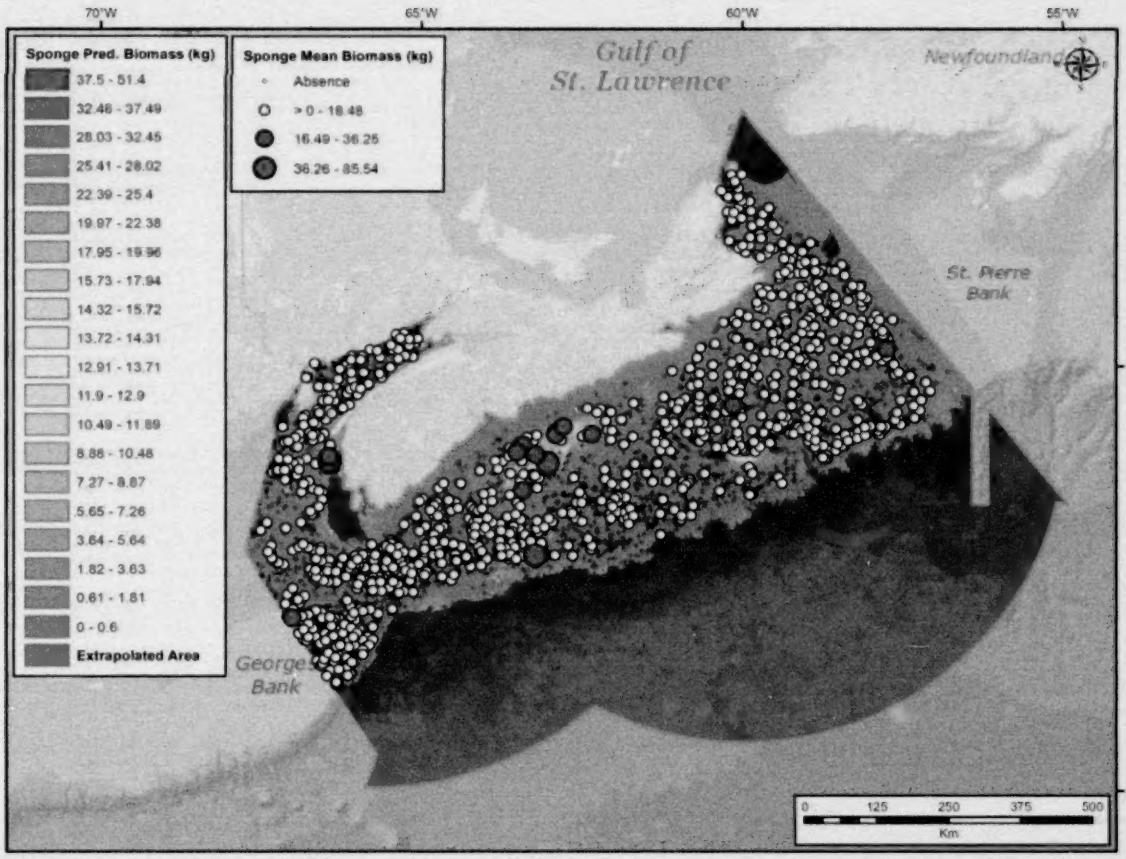

Figure 19. Predictions of biomass (kg) per grid cell of sponges from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2001 and 2015. Also shown are the mean biomass values per grid cell.

The top 15 most important environmental variables for predicting sponge biomass are shown in Figure 20. Summer Primary Production Mean was the most important variable in the model. Prior to spatial interpolation, this variable displayed a slightly bimodal distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed a spatial pattern to data points over- and under-predicted by a normal distribution, with over-predicted points located off the coast of Cape Breton and southwestern Nova Scotia and in the deepest regions of the study extent, and under-predicted points located in the centre of the study extent and in Emerald Basin. Summer Primary Production Mean was followed more distantly by Bottom Temperature Average Minimum and the other variables in the model. The partial dependence of sponge biomass on the top 6 most important variables is shown in Figure 21. Predicted biomass was highest at primary production values greater than 1100 mg C m^{-2} day^{-1}. Values in this range coincided with those data points under-predicted by a normal distribution off southwestern Nova Scotia. However, the fit between predicted and observed values for this variable was relatively

good, however, some points could be predicted lower than their true values and slightly outside the range of highest predicted biomass identified in the partial plot.

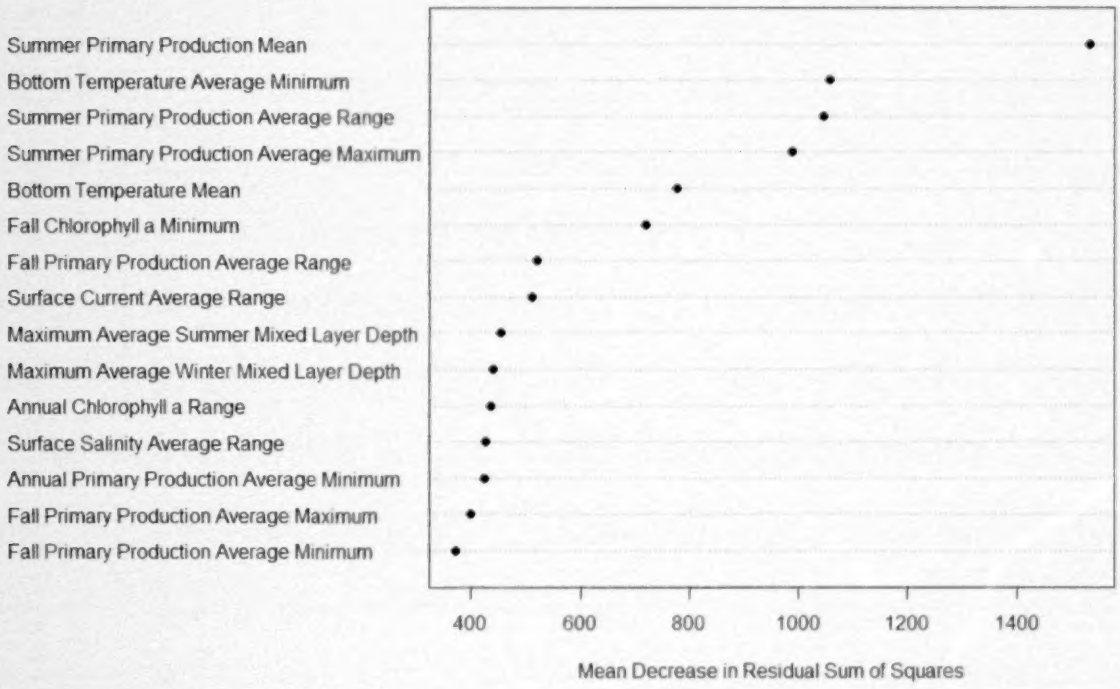

Figure 20. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the random forest model predicting sponge biomass (kg) within the Maritimes Region. The higher the Mean Decrease in Residual Sum of Squares the more important the variable is for predicting the response data.

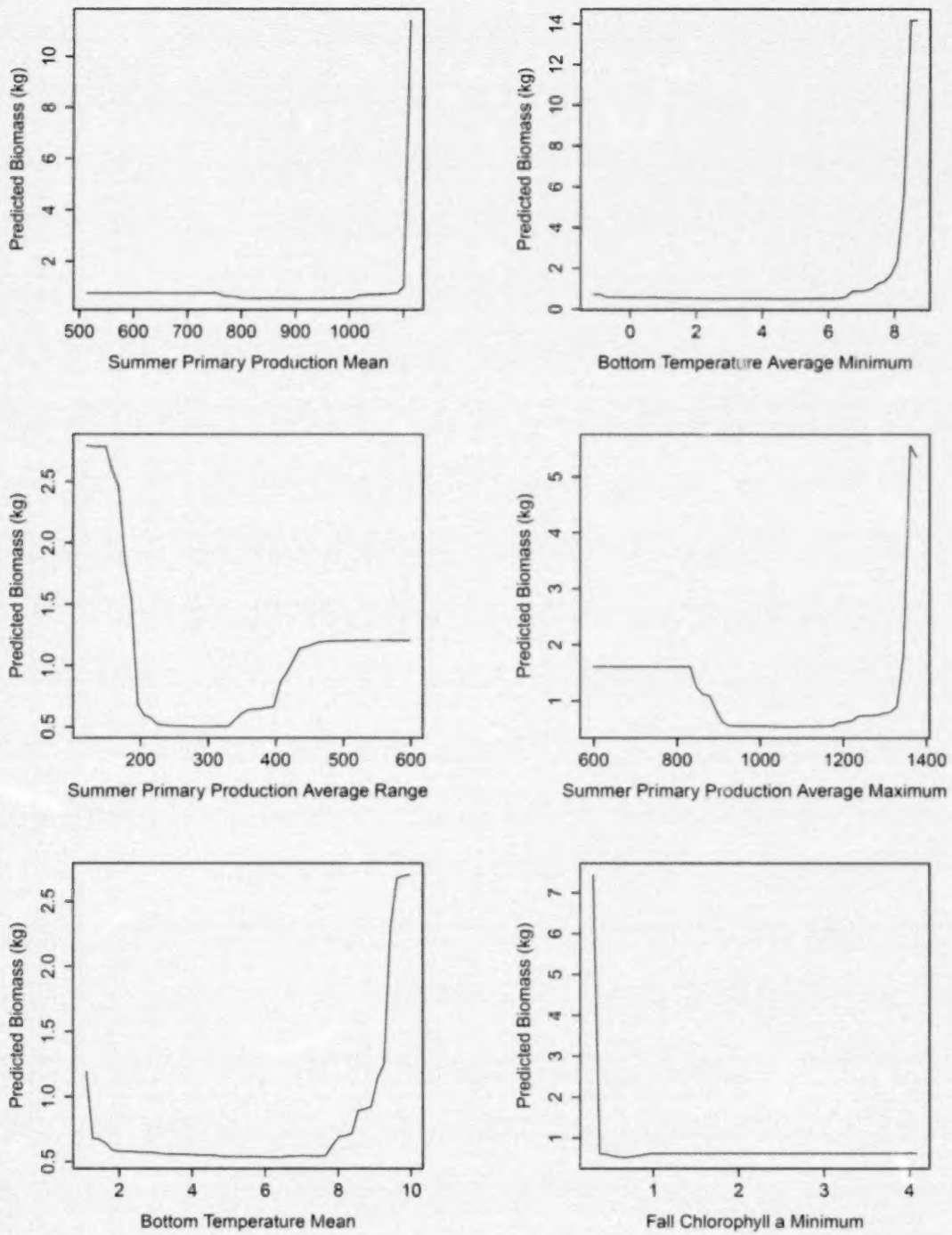

Figure 21. Partial dependence plots of the top six predictors from the optimal random forest model of sponge biomass collected within the Maritimes Region, ordered left to right from the top. Predicted biomass (kg) is shown on the *y*-axis.

Vazella pourtalesi (Russian Hat sponge)

Data Sources and Distribution

Figure 22 shows the distribution of available *Vazella pourtalesi* records in the Maritimes Region. Presence records of *V. pourtalesi* had an uneven spatial distribution across the study area, with the majority occurring in Emerald and LaHave Basins on central Scotian Shelf, and in deeper water between Emerald and LaHave Banks near the edge of the Scotian Shelf. There was a high degree of overlap in the spatial distribution of records originating from the different data sources. Records from multispecies and scientific surveys occurred in the Northeast Channel, while *V. pourtalesi* records from all three data sources occurred in the eastern Gulf of Maine.

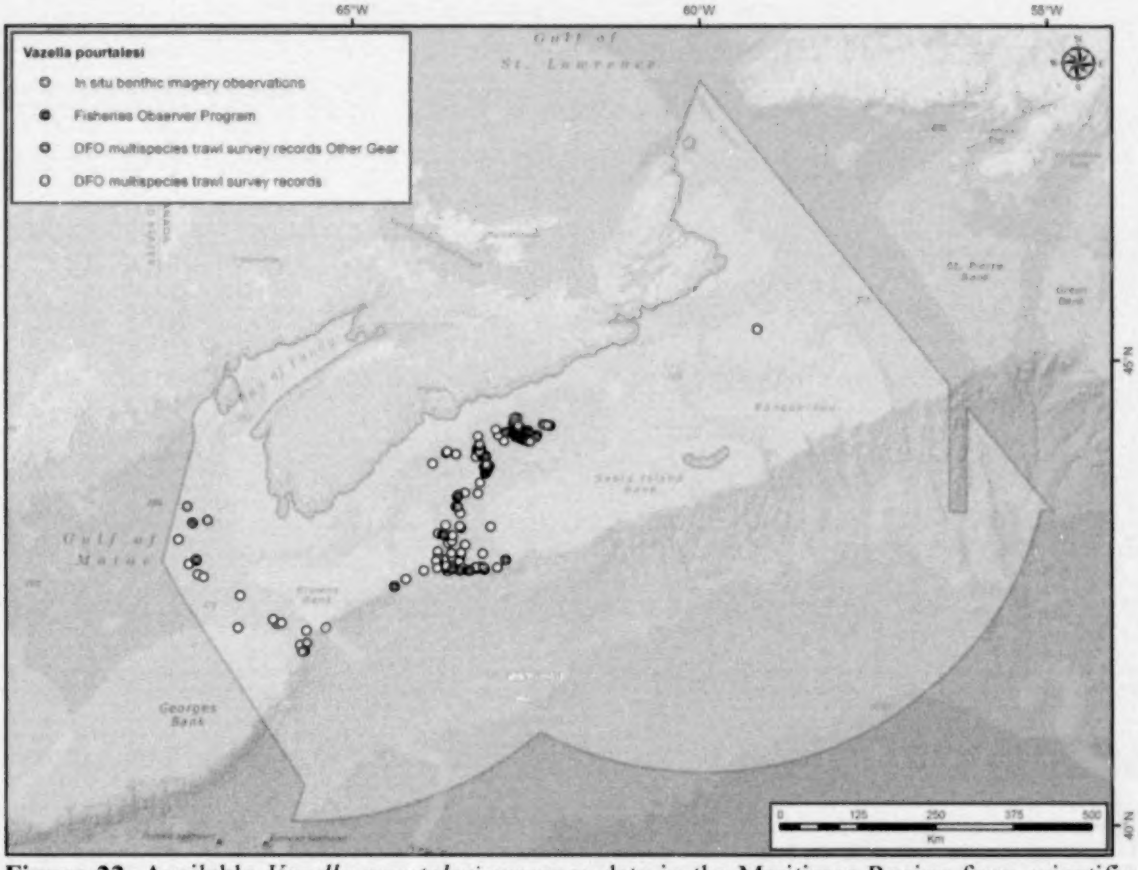

Figure 22. Available *Vazella pourtalesi* presence data in the Maritimes Region from scientific survey missions, DFO research vessel surveys, and commercial catches from the Fisheries Observer Program.

Initial random forest models of *Vazella pourtalesi* were run using only catch data originating from DFO multispecies trawl surveys (Western IIA gear). These records were collected over a period of 9 years from 2007 to 2015 (Table 7) and consisted of 60 presence and 1884 absences (Figure 23). Absence records were distributed relatively evenly across the Scotian Shelf and Bay of Fundy. The highest mean biomass record (84.54 kg) was recorded from Emerald Basin.

Table 7. Number of presence and absence, and total biomass of *Vazella pourtalesi* catch recorded from DFO multispecies trawl surveys between 2007 and 2015 conducted within the Maritimes Region.

Year	Total number of presences	Total number of absences
2007	4	173
2008	8	158
2009	1	198
2010	4	301
2011	8	251
2012	6	212
2013	10	296
2014	17	264
2015	2	31

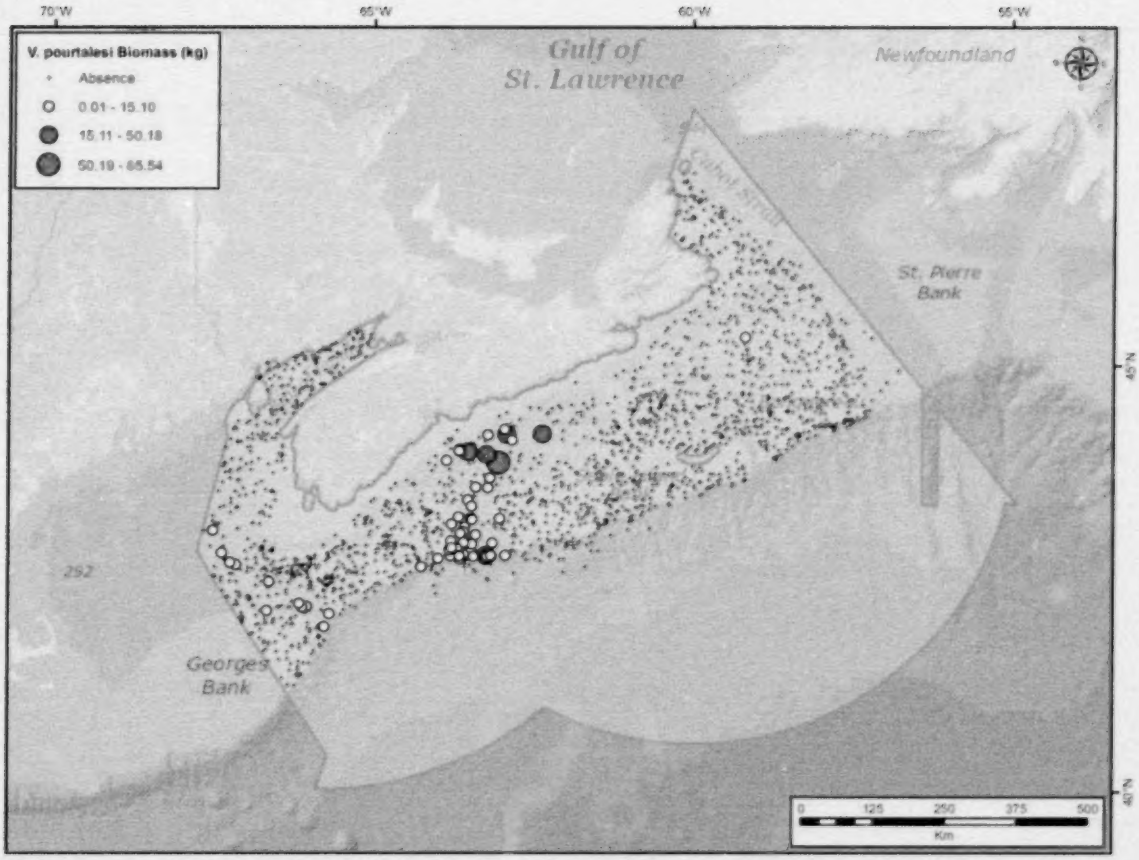

Figure 23. Biomass (kg) of *V. pourtalesi* catch recorded from DFO multispecies trawl surveys from 2007 to 2015 within the Maritimes Region.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity and specificity) for the random forest model on balanced species prevalence (60 presences and 60 absences; Model 1) are presented in Table 8. The highest AUC was 0.981 (Model Run 3). This model also had among the highest sensitivity and specificity measures of all 10 runs. The confusion matrix of this model is also presented in Table 8. Class error for both the presence and absence classes was low (0.050 and 0.100, respectively).

The presence probability prediction surface of *V. pourtalesi* from Model 1 is presented in Figure 24. The highest predictions of presence probability occurred in Emerald Basin, LaHave Basin, and the area between LaHave Bank and Emerald Bank near the shelf break. The Northeast Channel and eastern Gulf of Maine also had a high probability of presence of *V. pourtalesi*. These areas of high presence probability corresponded well with the spatial distribution of

presence records (see Figure 25). Presence probability of *V. pourtalesi* was moderate to high along the shelf break despite the absence of presence records there.

Table 8. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of *V. pourtalesi* within the Maritimes Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 3) which is considered the optimal model for predicting the presence probability of *V. pourtalesi*.

Model Run	AUC	Sensitivity	Specificity
1	0.940	0.867	0.833
2	0.932	0.883	0.817
3	**0.981**	**0.933**	**0.883**
4	0.962	0.933	0.867
5	0.954	0.933	0.900
6	0.962	0.917	0.833
7	0.917	0.833	0.783
8	0.946	0.917	0.883
9	0.927	0.917	0.817
10	0.952	0.900	0.850
Mean	**0.947**	**0.903**	**0.847**
SD	**0.019**	**0.033**	**0.037**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	54	6	60	0.100
Presence	3	57	60	0.050

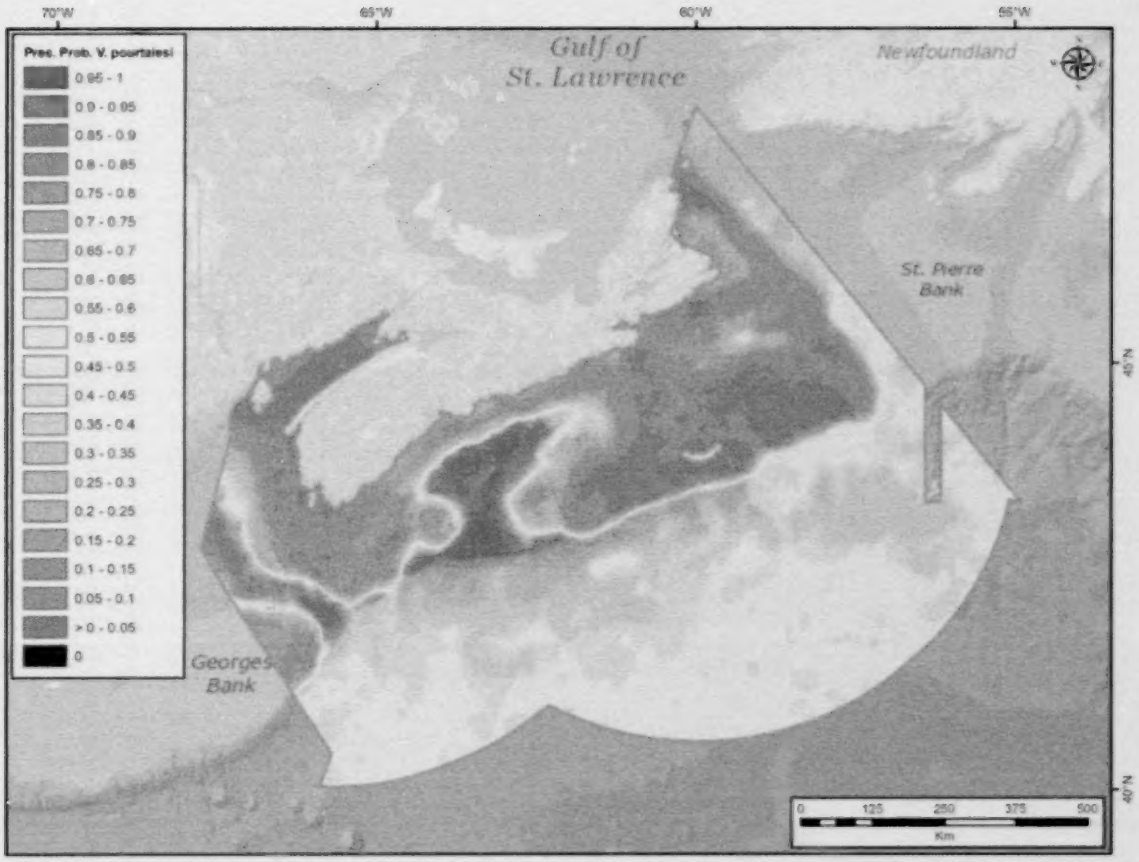

Figure 24. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of *Vazella pourtalesi* presence and absence data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2007 and 2015.

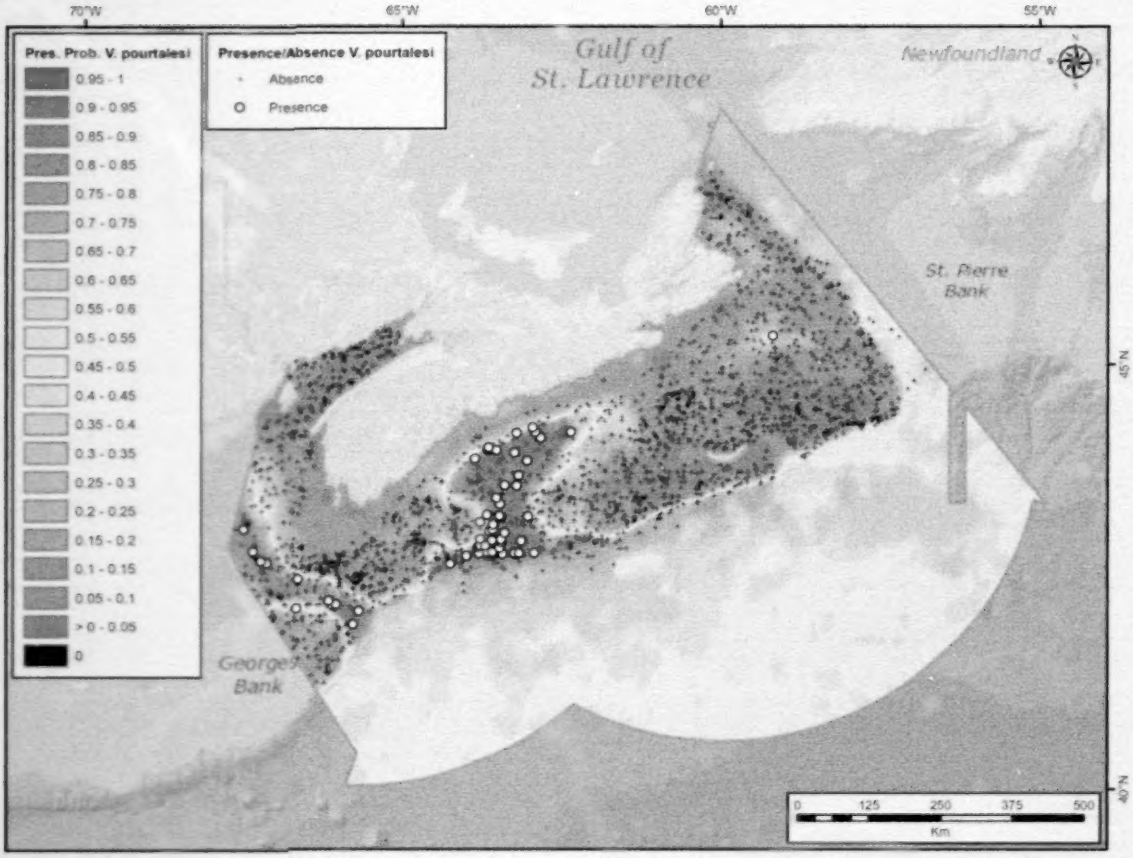

Figure 25. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of *Vazella pourtalesi* presence and absence data recorded from DFO multispecies trawl surveys conducted within the Maritimes Region between 2007 and 2015.

The actual presence and absence data observations used in the final model run of Model 1 (60 presences and 60 absences; Figure 26) showed extreme spatial bias across the study area. Despite their being absence records in Emerald Basin, LaHave Basin, and Northeast Channel, only a few absence records were selected from these areas during the random down-sampling of the data prior to modelling. This likely caused the over-extension of high predicted probabilities in these areas beyond where presence data occurred. Areas of extrapolation of Model 1 are also shown in Figure 26. Several small pockets of extrapolated area overlap with areas of high predicted presence probability of *V. pourtalesi* in Emerald Basin, the Northeast Channel, and eastern Gulf of Maine.

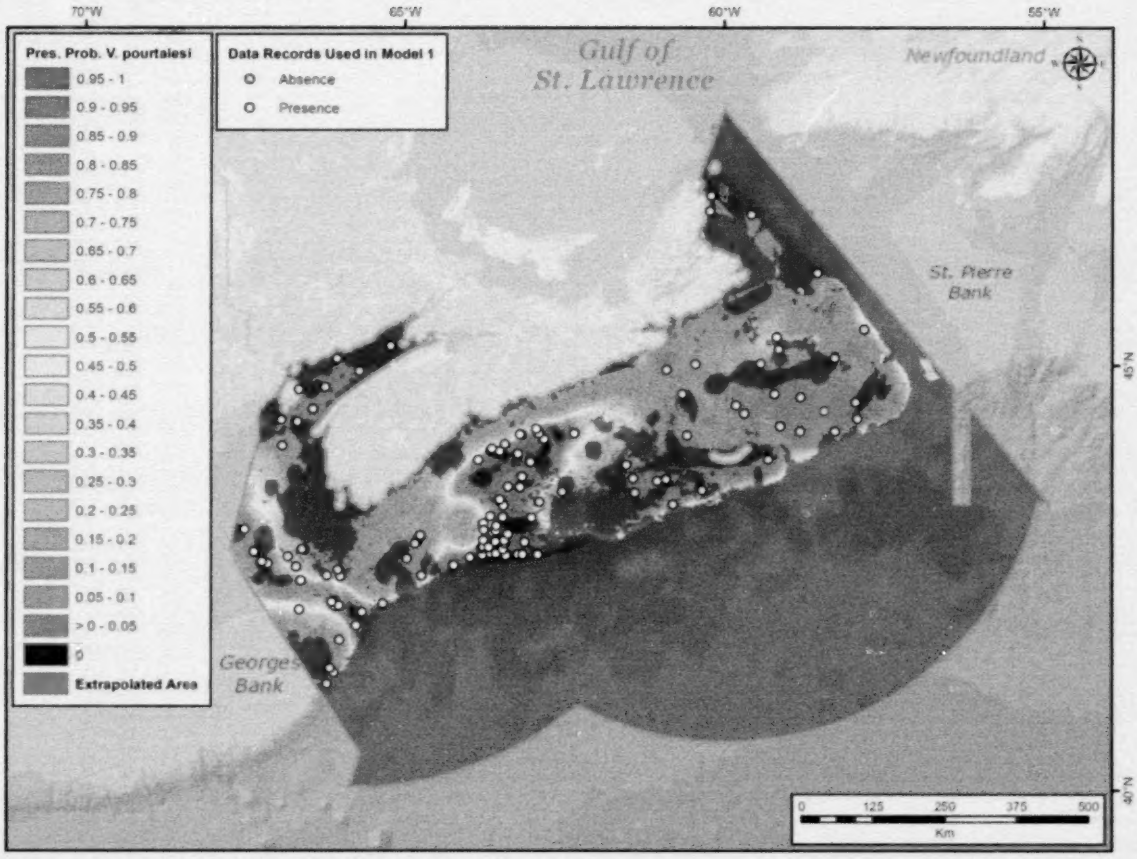

Figure 26. Map of the 120 data observations (60 presences and 60 absences) of *V. pourtalesi* used in the optimal random forest Model 1. Also shown is the predicted presence probability (Pres. Prob.) of *V. pourtalesi* generated from Model 1 and areas of model extrapolation.

Of all 66 environmental predictor variables used in the model, Bottom Salinity Average Minimum was the most important for the classification of *V. pourtalesi* response data (Figure 27). Prior to spatial interpolation, this variable displayed a left-skewed distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed a relatively weak spatial pattern to data points over- and under-predicted by a normal distribution, with over-predicted points located in the Laurentian Channel, Gulf of Maine, central Scotian Shelf, and in the deepest regions of the study extent, and under-predicted points located around the coast and near the edge of the shelf. This variable was followed closely in terms of its Mean Decrease in Gini Value by Bottom Salinity Average Maximum and Bottom Salinity Mean. Depth was the 6th most important variable in the model. Partial dependence plots for the top 6 predictor variables are shown in Figure 28. In general, presence probability of *V. pourtalesi* was highest at salinity values greater than 34. Values in this range coincided with both over-predicted data points in Emerald Basin and under-predicted data points in the Northeast Channel. The fit between predicted and observed values

for this variable was fair, with some deviation in data points from the 1:1 reference line with slight under-prediction of salinity values 34 and greater. However, these values were still within the range of high presence probability identified in the partial plot (Figure 28). Along the Depth gradient, presence probability of *V. pourtalesi* showed a sharp increase at approximately ~150 m depth.

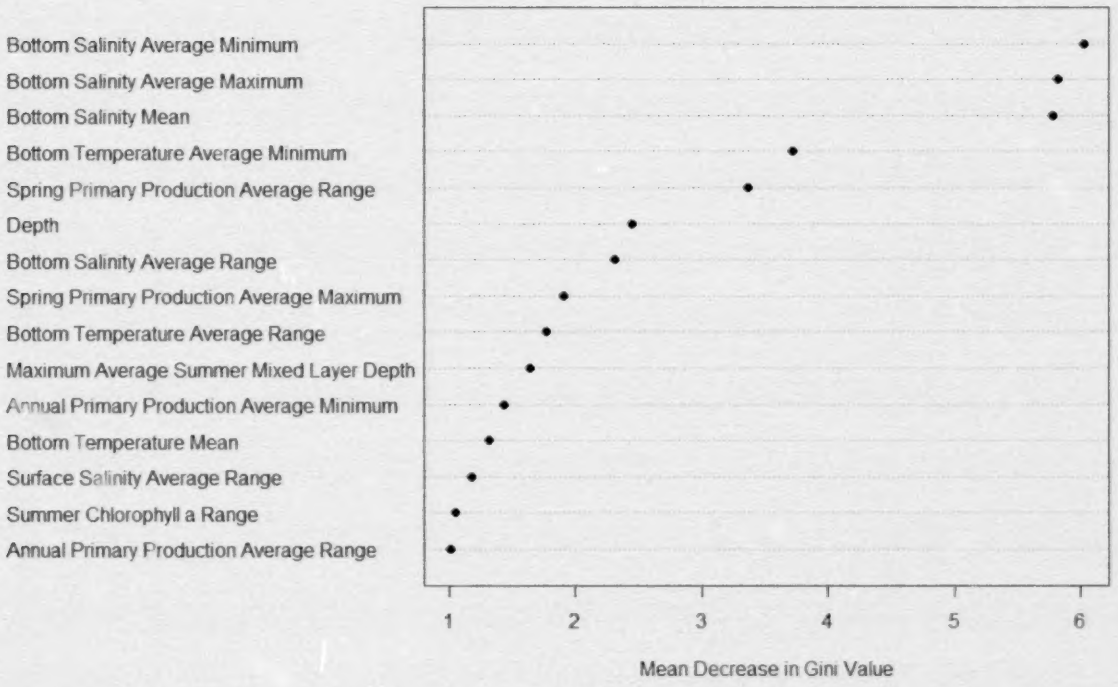

Figure 27. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting *Vazella pourtalesi* presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

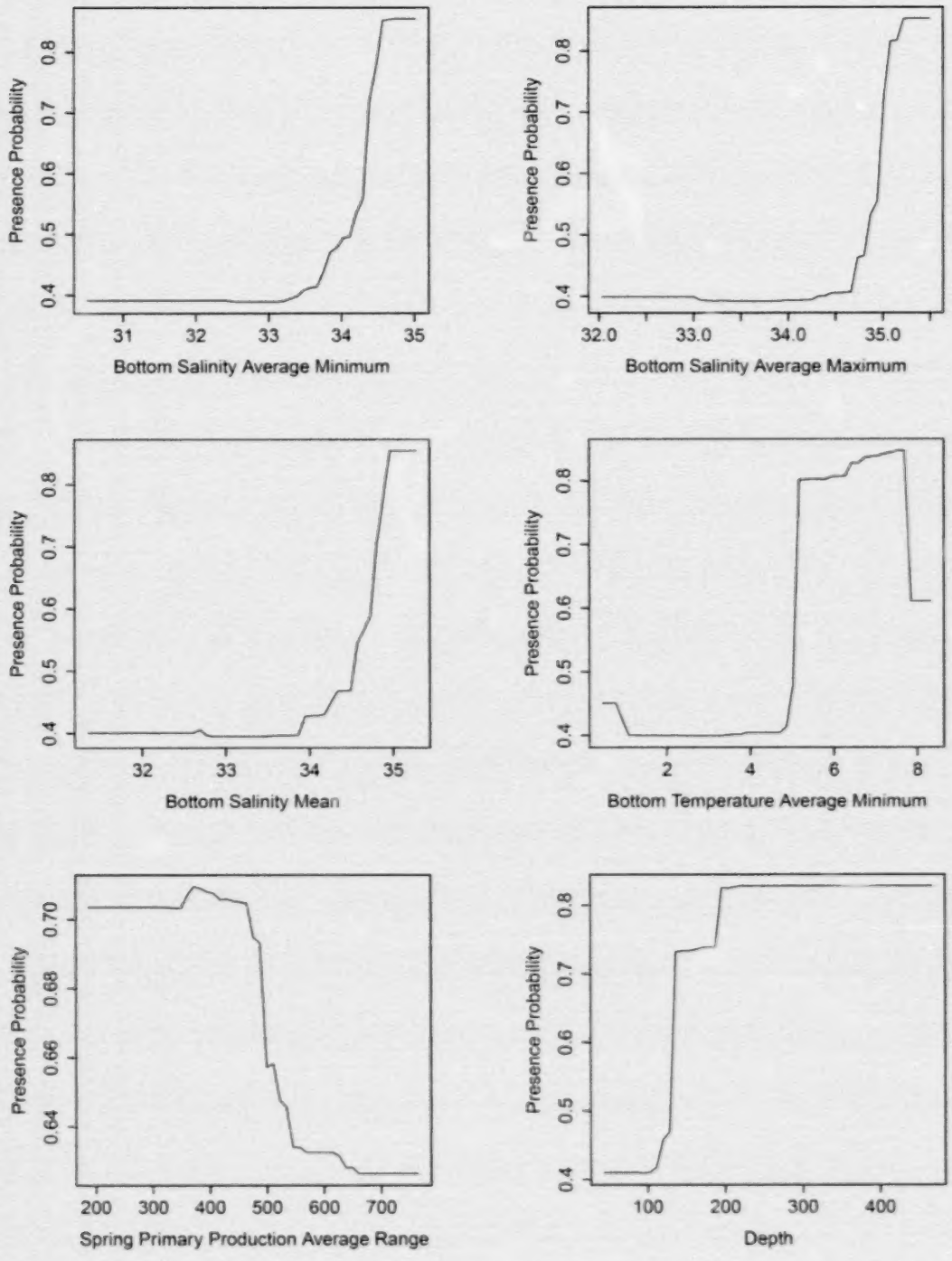

Figure 28. Partial dependence plots of the top six predictors from the optimal random forest model of *Vazella pourtalesi* presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 9 shows accuracy measures for the random forest model on all *V. pourtalesi* presence and absence records and a threshold equal to species prevalence (0.03). The average AUC computed from 10-fold cross validation was 0.930, slightly lower than the average AUC from Model 1. Class error for the presence class was comparable to Model 1, whereas error for the absence class was slightly higher than Model 1. Sensitivity and specificity measures were similar between both models.

Predicted probabilities of presence of *V. pourtalesi* generated from Model 2 were much more conservative than that of Model 1 (Figure 29). Presence probability was high in small, isolated areas in Emerald Basin, along the shelf break between LaHave and Emerald Banks, and in the Northeast Channel. The majority of the eastern Scotian Shelf and Bay of Fundy where only absence records occur were predicted to have zero presence of *V. pourtalesi* (Figure 30). Areas of extrapolation are shown in Figure 31.

Table 9. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of *V. pourtalesi* within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.966		**Absence**	**Presence**				
2	0.987	**Absence**	1606	278	1884	0.148	0.917	0.852
3	0.985	**Presence**	5	55	60	0.083		
4	0.935							
5	0.857							
6	0.993							
7	0.941							
8	0.946							
9	0.981							
10	0.711							
Mean	**0.930**							
SD	**0.087**							

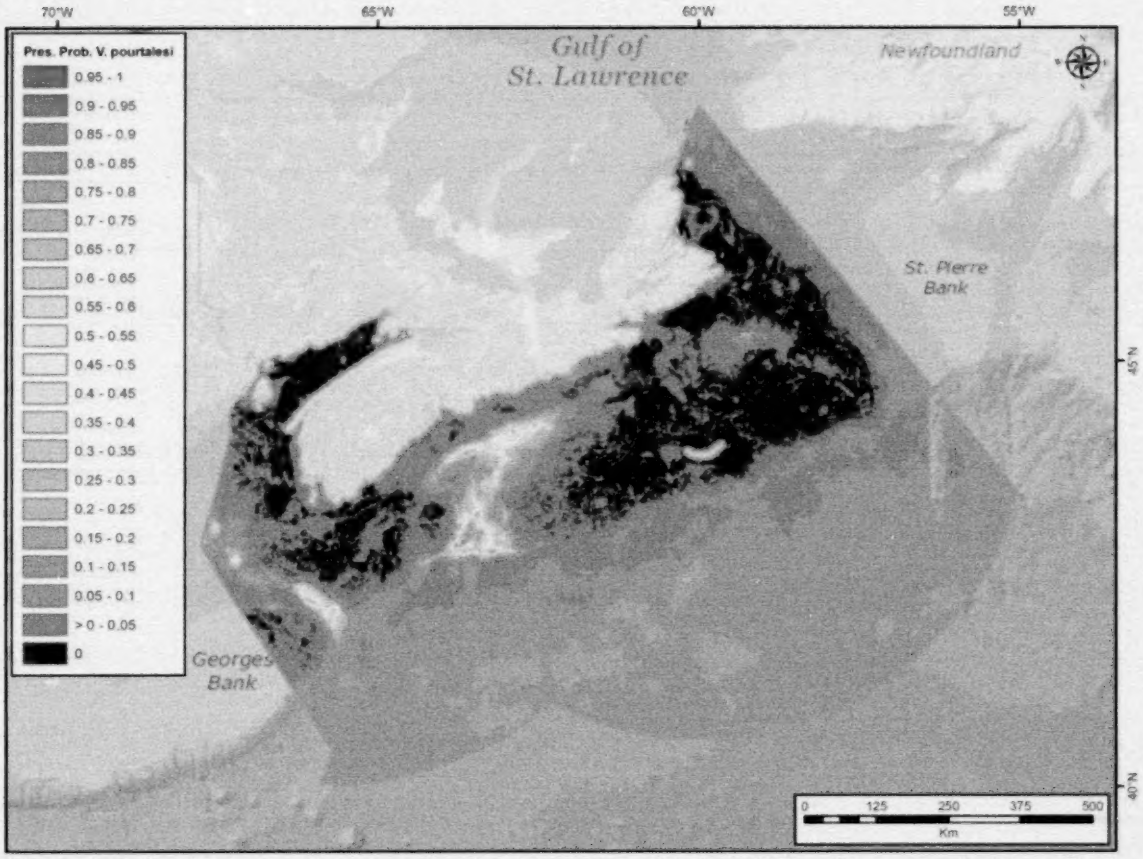

Figure 29. Predictions of presence probability (Pres. Prob.) of *Vazella pourtalesi* based on a random forest model on unbalanced presence and absence *V. pourtalesi* catch data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2007 and 2015.

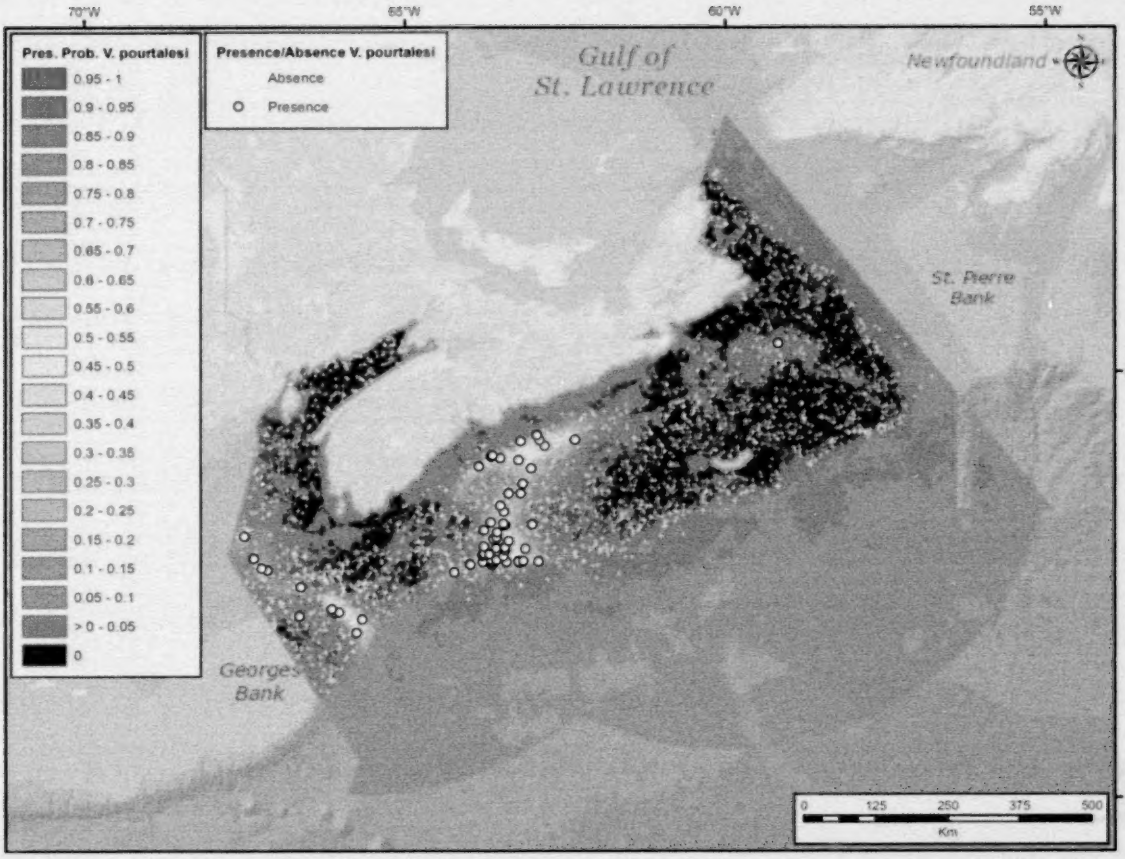

Figure 30. Presence and absence observations and predictions of presence probability (Pres. Prob.) of *Vazella pourtalesi* based on a random forest model on unbalanced presence and absence *V. pourtalesi* catch data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2007 and 2015.

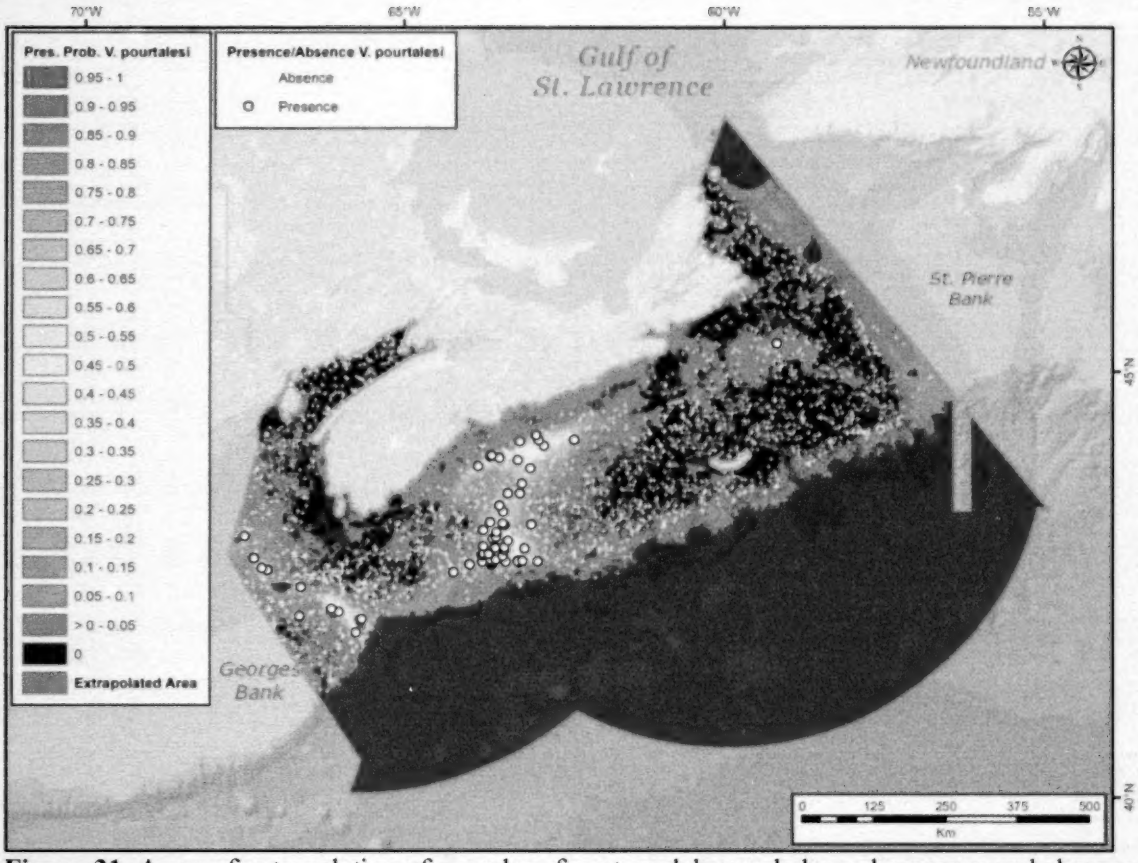

Figure 31. Areas of extrapolation of a random forest model on unbalanced presence and absence *V. pourtalesi* catch data collected within the Maritimes Region between 2007 and 2015. Also shown are the presence and absence observations and predictions of presence probability (Pres. Prob.).

The order of importance of the environmental predictor variables in Model 2 (Figure 32) was slightly different than that of Model 1. Bottom Salinity Average Maximum was the most important variable in Model 2 compared to Bottom Salinity Average Minimum in Model 1. Bottom Salinity Average Maximum displayed a left-skewed distribution prior to spatial interpolation (Beazley et al., in prep). Examination of the Q-Q plot revealed a spatial pattern to data points over- and under-predicted by a normal distribution, with over-predicted points located in Emerald Basin, northern Laurentian Channel, and in the deepest regions of the study extent, and under-predicted points also located in Emerald Basin and over much of central Scotian Shelf. This variable was followed more distantly by Bottom Salinity Mean and Bottom Temperature Average Minimum. Partial dependence plots for the top 6 predictor variables are

shown in Figure 33. In general, presence probability of *V. pourtalesi* was highest at Bottom Salinity Average Maximum values greater than 35. Values in this range coincided with under-predicted data points in Emerald Basin and along central Scotian Shelf. The fit between predicted and observed values for this variable was fair, with some deviation in data points from the 1:1 reference line with slight under-prediction of salinity values 35 and greater. However, these values were still within the range of high presence probability identified in the partial plot (Figure 33).

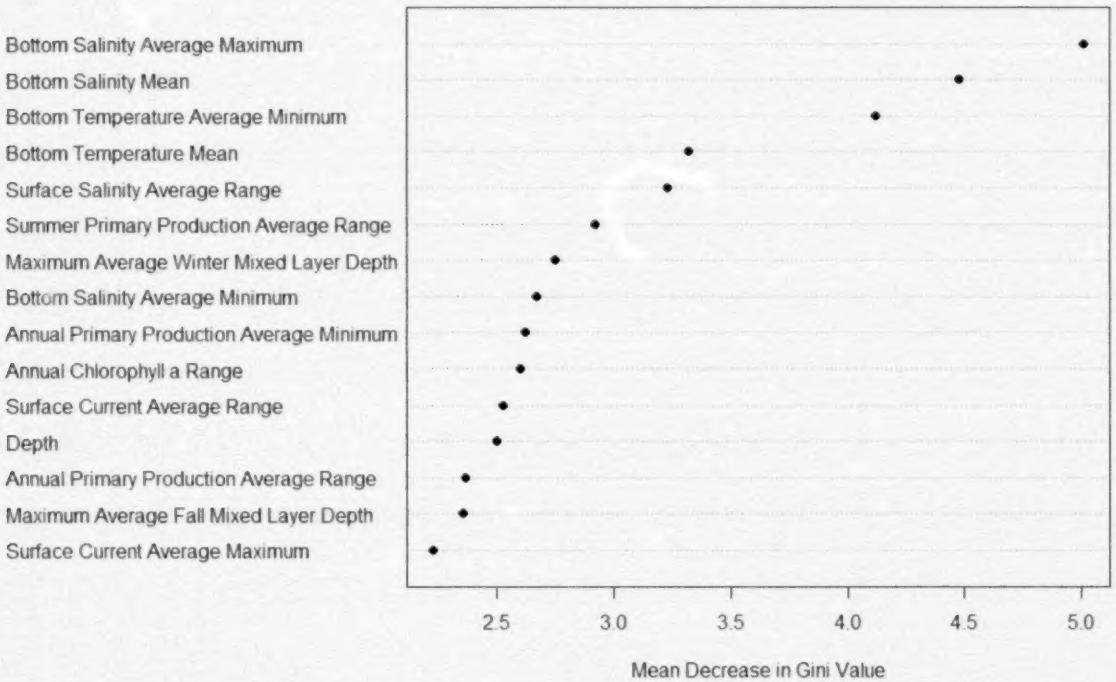

Figure 32. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced *Vazella pourtalesi* presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

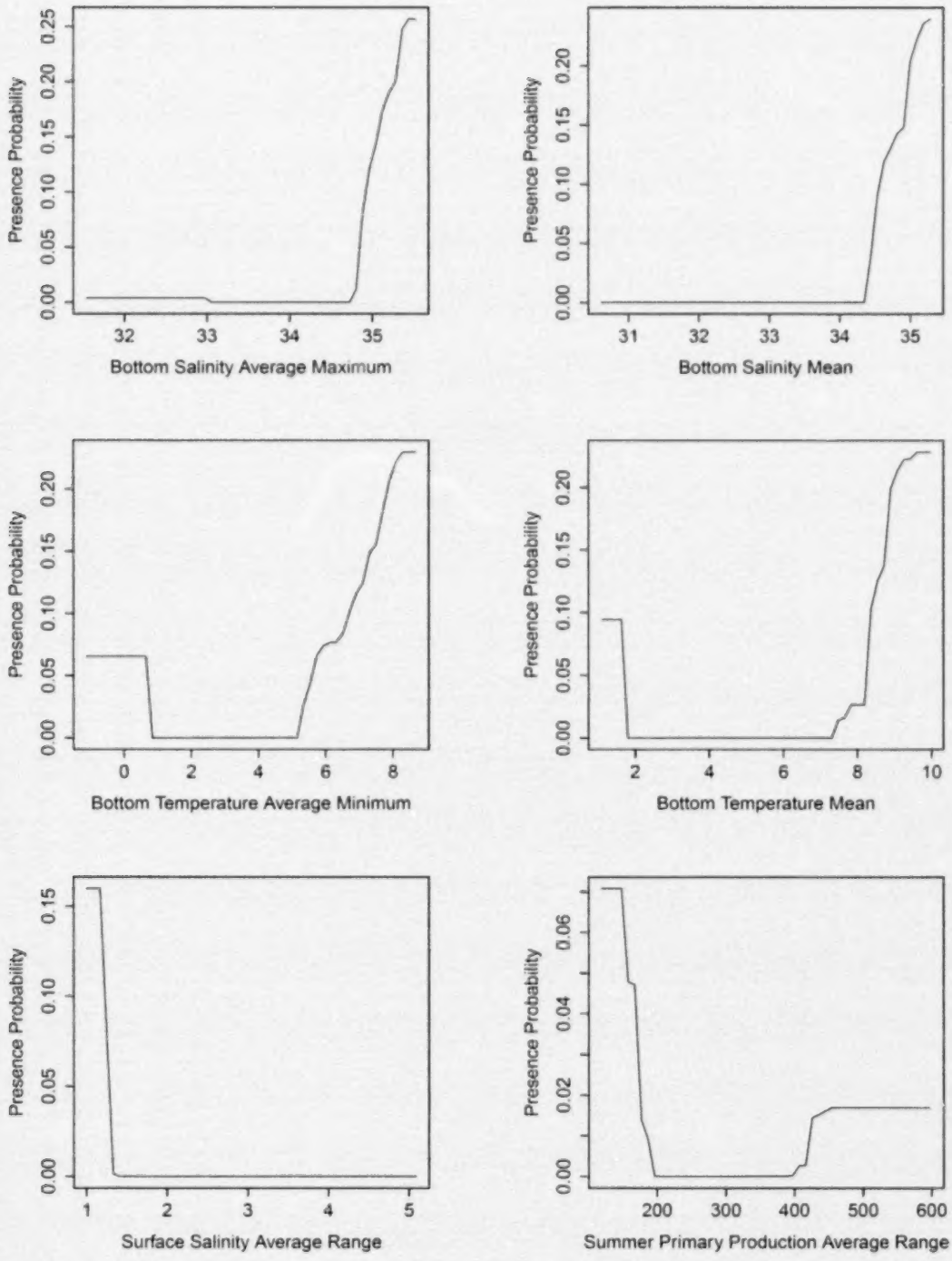

Figure 33. Partial dependence plots of the top six predictors from the random forest model of *Vazella pourtalesi* unbalanced presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 3 – Addition of Commercial Records and *In Situ* Benthic Imagery Observations

Given the low number of presence records of this unique population of *V. pourtalesi*, the DFO multispecies trawl survey data were augmented with presence records from all available data sources (see Figure 22 and Table 10). The combined dataset, consisting of 166 presences and 1983 absences, was remodelled (termed Model 3) using an unbalanced design and a threshold equal to species prevalence (0.08).

Accuracy measures for Model 3 are shown in Table 11. The average AUC computed from 10-fold cross validation was 0.977, the highest of all three models. Class error for the presence and absence classes is comparable to Model 1. Sensitivity and specificity measures were both high.

The additional presence records expanded the area of high presence probability in Emerald and LaHave Basins compared to Model 2 (Figure 34). Similarly, parts of the Northeast Channel also have a higher probability of presence of *V. pourtalesi*. The majority of the eastern Scotian Shelf and Bay of Fundy where only absence records occurred were predicted to have zero presence of *V. pourtalesi* (Figure 35). Areas of extrapolation were similar to that of Model 2 (Figure 36). Figure 37 depicts the classification of *V. pourtalesi* presence probability into presence and absence categories based on the prevalence threshold of 0.08. Part of the Gulf of Maine, Northeast Channel, Emerald and LaHave Basins, and the Laurentian Channel were classified as presence of *V. pourtalesi*.

Table 10. Additional presence records of *Vazella pourtalesi* from *in situ* benthic surveys, the Fisheries Observer Program, and DFO multispecies trawls using US 4 seam 3 bridle gear collected within the Maritimes region between 1997 and 2015. FOP = Fisheries Observer Program.

Mission	Year	Gear	Total number of presences
HUD2005-186	2005	Campod	3
HUD2011-014	2011	Campod	23
2001ROPOS	2001	ROPOS	3
FOP	1997 to 2007, and 2011-2015	Commercial trawl gear	75
NED2013	2013	US 4 seam 3 bridle trawl	1
NED2014	2014	US 4 seam 3 bridle trawl	1

Table 11. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of *V. pourtalesi* from DFO multispecies trawl surveys, the Fisheries Observer Program, and *in situ* benthic imagery observations collected within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.978		**Absence**	**Presence**				
2	0.990	**Absence**	1811	172	1983	0.087	0.952	0.913
3	0.948	**Presence**	8	158	166	0.048		
4	0.980							
5	0.978							
6	0.983							
7	0.963							
8	0.978							
9	0.980							
10	0.993							
Mean	**0.977**							
SD	**0.013**							

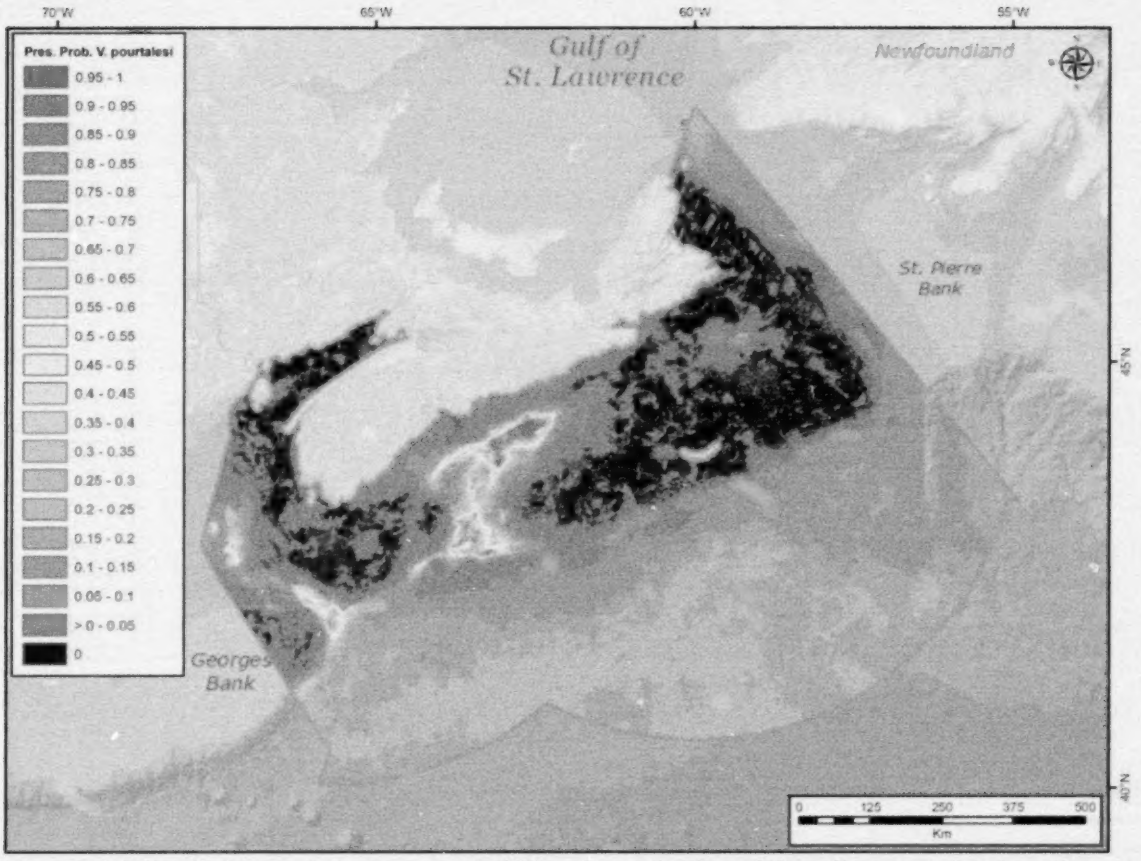

Figure 34. Predictions of presence probability (Pres. Prob.) of *Vazella pourtalesi* based on a random forest model on unbalanced presence and absence *V. pourtalesi* data collected from DFO multispecies trawl surveys, the Fisheries Observer Program, and scientific surveys conducted within the Maritimes Region between 1997 and 2015.

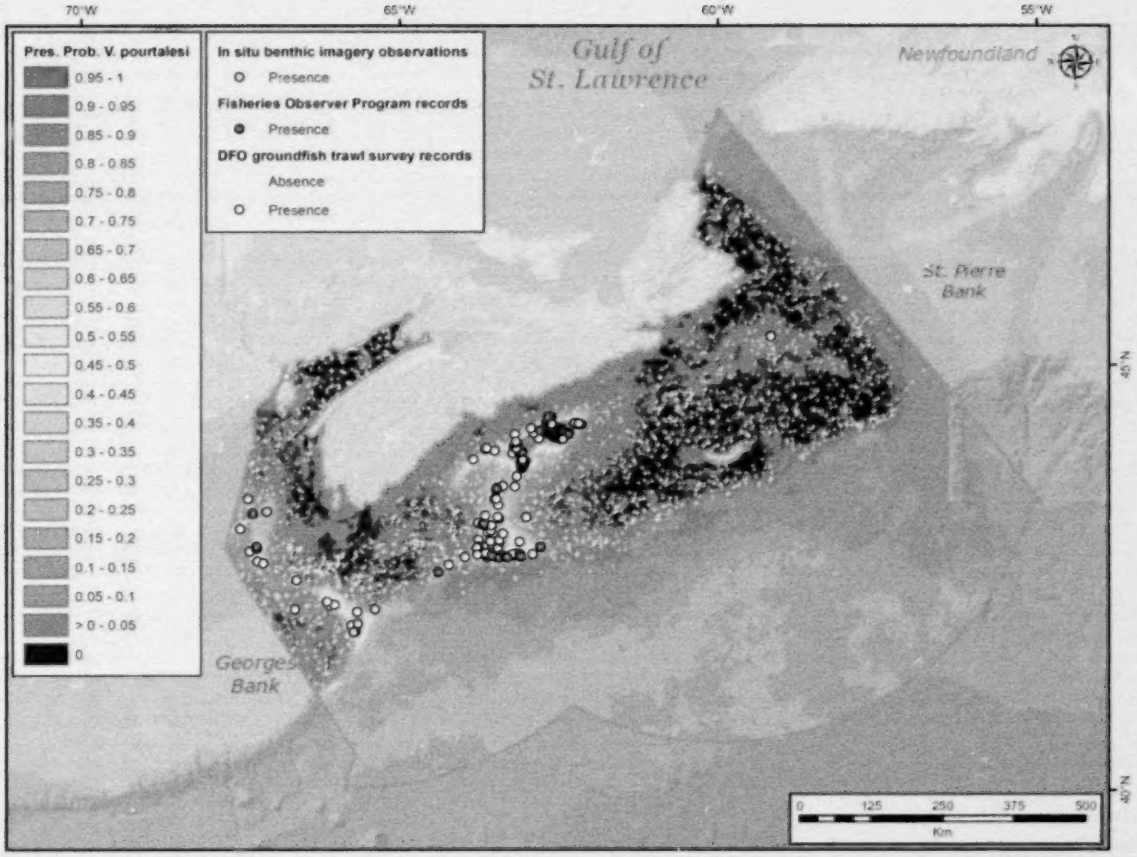

Figure 35. Presence and absence observations and predictions of presence probability (Pres. Prob.) of *Vazella pourtalesi* based on a random forest model on unbalanced presence and absence *V. pourtalesi* data collected from DFO multispecies trawl surveys, the Fisheries Observer Program, and scientific surveys conducted within the Maritimes Region between 1997 and 2015.

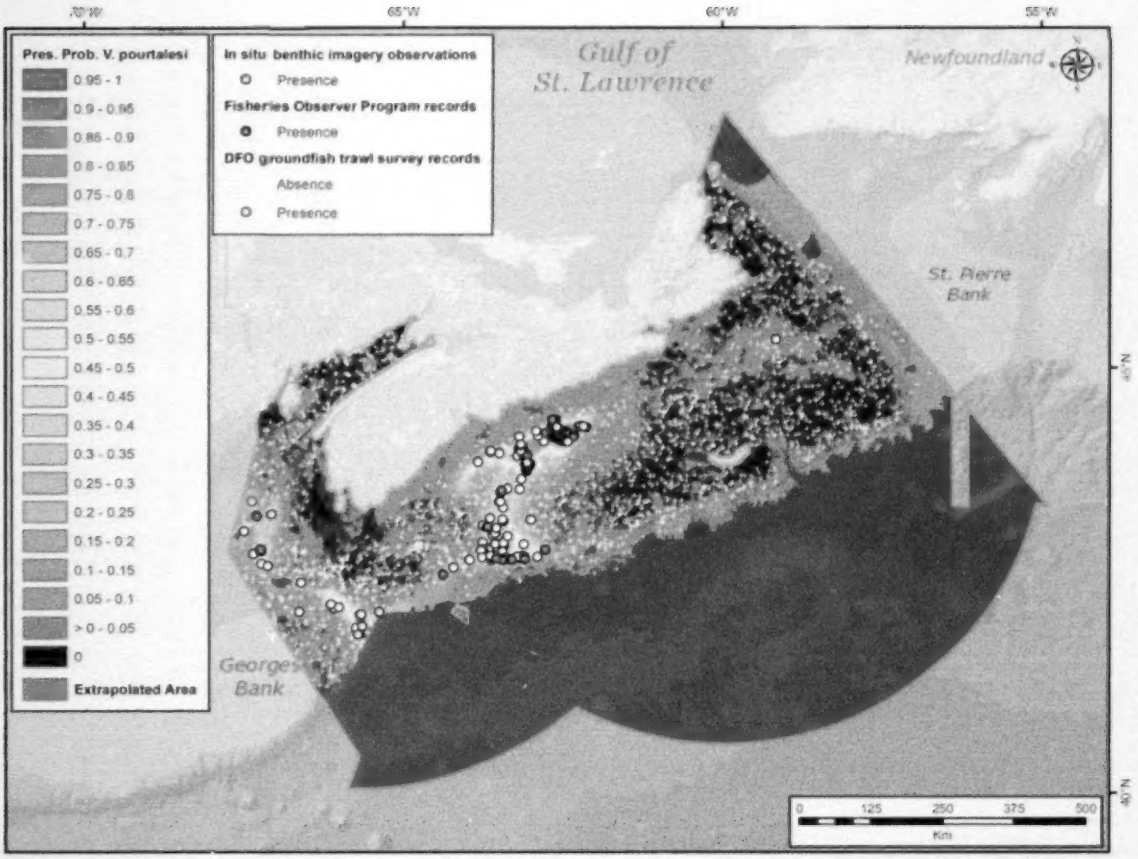

Figure 36. Areas of extrapolation of Model 3 on unbalanced presence and absence *V. pourtalesi* data collected from DFO multispecies trawl surveys, the Fisheries Observer Program, and scientific surveys conducted within the Maritimes Region between 1997 and 2015. Also shown are the presence and absence observations and predictions of presence probability (Pres. Prob.).

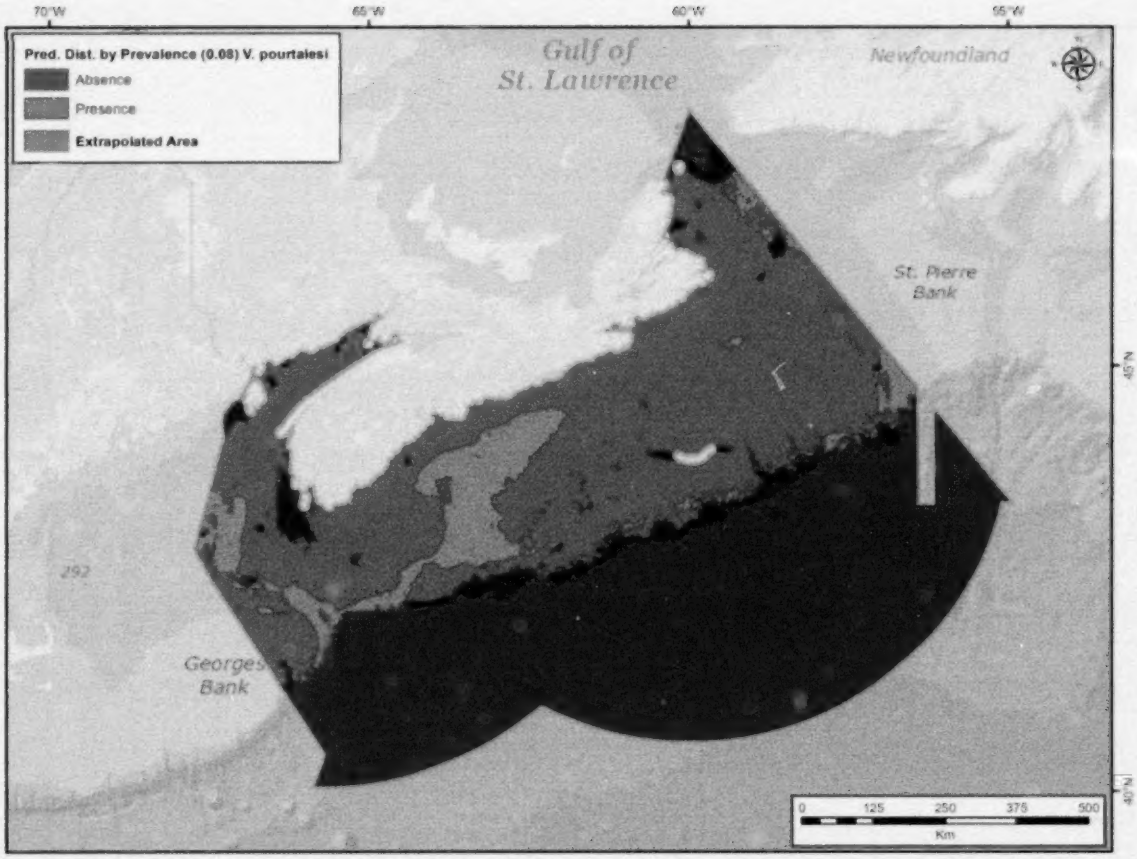

Figure 37. Predicted distribution (Pred. Dist.) of *Vazella pourtalesi* in the Maritimes Region based on the prevalence threshold of 0.08 of *V. pourtalesi* presence and absence data used in Model 3. Also shown are the areas of model extrapolation (grey polygon may appear red or blue).

The order of importance of environmental predictor variables in this model (Figure 38) was similar to that of Model 2. As in Model 2, Bottom Salinity Average Maximum was the most important variable, followed by Bottom Salinity Mean, and more distantly by Bottom Temperature Average Minimum. Like Model 2, several primary production variables held higher positions in terms importance in the model compared to Model 1. Partial dependence plots are shown in Figure 39. Similar to Models 1 and 2, presence probability of *V. pourtalesi* was highest at the highest salinity and temperature values.

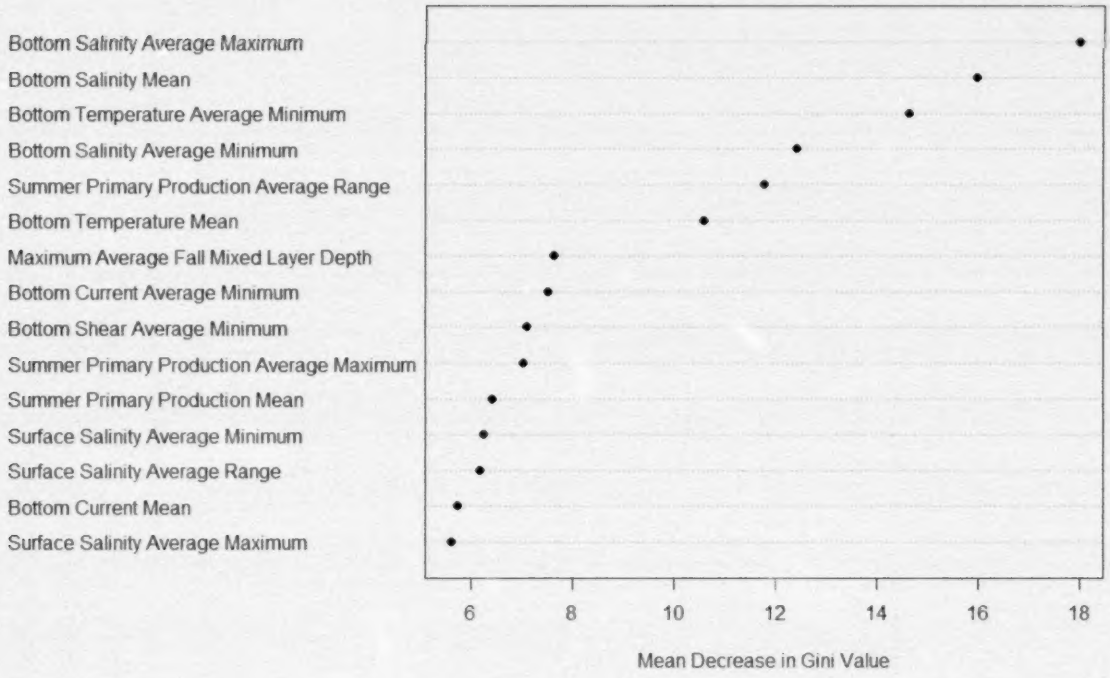

Figure 38. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced *Vazella pourtalesi* presence and absence data collected from DFO multispecies trawl surveys, the Fisheries Observer Program, and scientific surveys conducted within the Maritimes Region between 1997 and 2015. The higher the Mean Gini value the more important the variable is for predicting the response data.

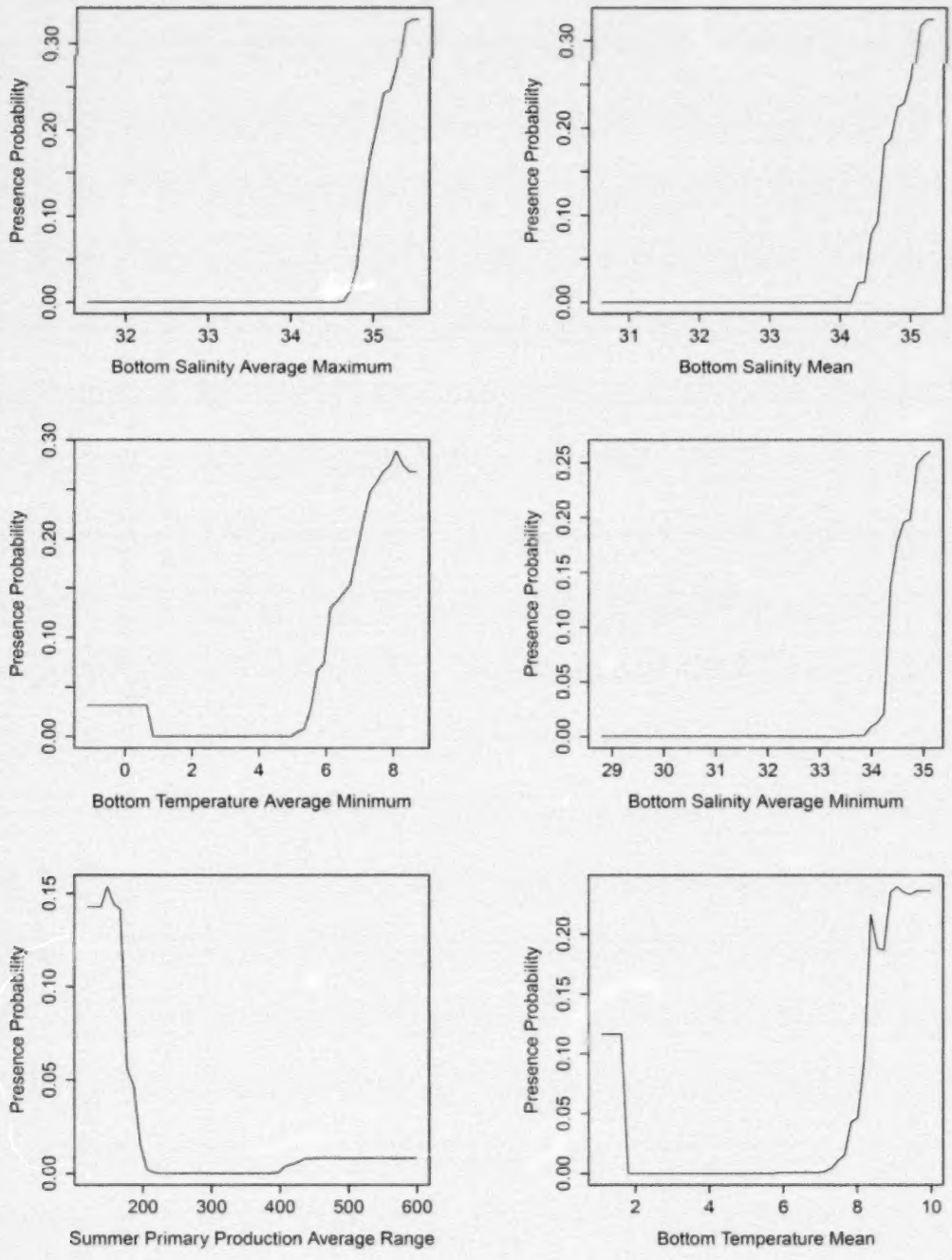

Figure 39. Partial dependence plots of the top six predictors from the random forest model of *Vazella pourtalesi* unbalanced presence and absence data collected from DFO multispecies trawl surveys, the Fisheries Observer Program, and scientific surveys conducted within the Maritimes Region between 1997 and 2015, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all available *V. pourtalesi* records and unbalanced species prevalence (Model 3) was selected as the best predictor of the distribution of this species in the Maritimes Region. Model 1 (balanced species prevalence) was considered a poor predictor of presence probability of *V. pourtalesi* due to its exaggerated high presence probability beyond the location of presence data in Emerald and LaHave Basins. This phenomenon was likely due to random down-sampling of the absence data. Model 2, which was generated using the same presence-absence dataset but using all absence data, produced a much more realistic presence probability surface with less exaggeration beyond the location of presence points. The additional presence records added to Model 3 produced the highest AUC and sensitivity and specificity measures of all three models. Although the presence probability surface was similar to that of Model 2, areas of high presence probability were expanded in Emerald and LaHave basins due to the additional presence records, providing a more accurate depiction of the distribution of this species in that area. Note that there were no additional records of *V. pourtalesi* for use in model validation. All available records of this species were used in Model 3.

Figure 40 depicts the predictions of presence probability of *V. pourtalesi* from Model 3 in Emerald Basin in relation to DFO's Emerald Bank and Sambro Bank *Vazella* Closures. Although

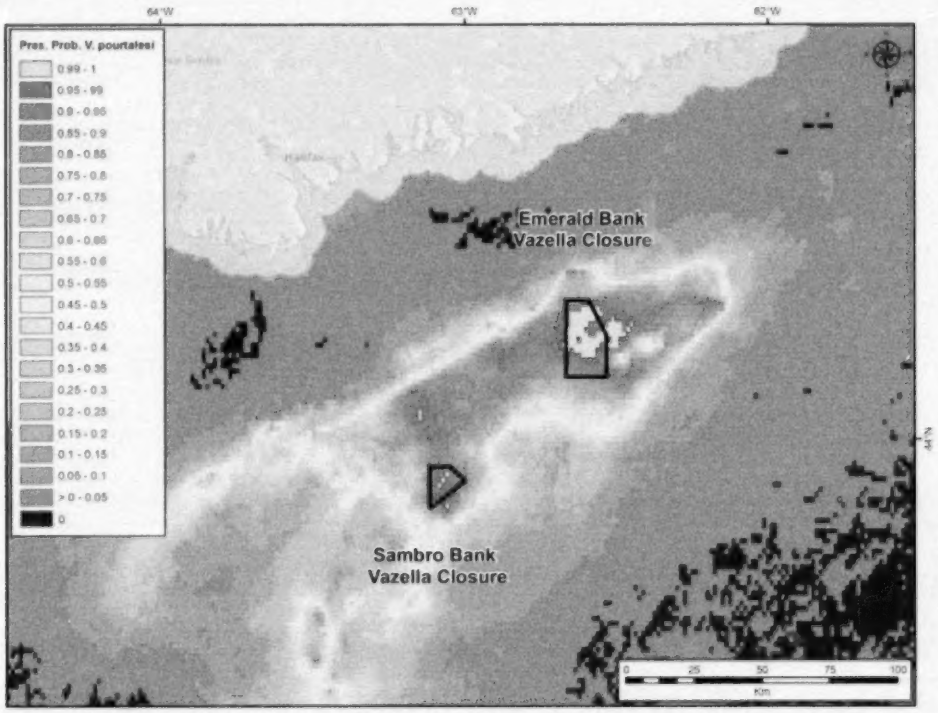

Figure 40. Predictions of presence probability of *Vazella pourtalesi* in Emerald Basin in relation to the location of DFO's Emerald Bank and Sambro Bank *Vazella* Closures.

the Emerald Bank Closure encompasses much of the area with the highest predicted probability of occurrence of *V. pourtalesi* (99 to 100% occurrence; green area in Figure 40), much of the area in Emerald Basin outside the two closures was predicted to have a relatively high probability of occurrence of *V. pourtalesi*. When considering species prevalence (see Figure 37), all of Emerald Basin is considered suitable habitat for this species.

Prediction of *Vazella pourtalesi* Biomass using Random Forest

The accuracy measures of the regression random forest model on mean *V. pourtalesi* biomass per grid cell from DFO multispecies trawl surveys are presented in Table 12. The highest R^2 was 0.207, while the average was 0.087 ± 0.079 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.024 ± 0.021 SD. The high standard deviation values for both of these metrics indicate high variability between model folds. The highest percent variance explained was 1.16%. The majority of the model folds had a negative variance explained, indicating poor predictive performance of the model.

Table 12. Accuracy measures from 10-fold cross validation of a random forest model of *Vazella pourtalesi* biomass (kg) per grid cell recorded from DFO multispecies trawl surveys in the Maritimes Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.192	6.913	0.081	-7.07
2	0.099	0.703	0.008	-2.26
3	0.129	2.175	0.025	-6.54
4	0.032	2.461	0.029	1.16
5	0.155	1.960	0.023	-4.41
6	2.860×10^{-4}	1.962	0.023	0.72
7	0.028	0.838	0.010	-1.96
8	0.207	2.218	0.026	-1.05
9	0.028	1.105	0.013	-1.84
10	0.004	0.563	0.007	-4.47
Mean	**0.087**	**2.090**	**0.024**	**-2.77**
SD	**0.079**	**1.833**	**0.021**	**2.80**

Figures 41 and 42 show the predicted biomass surface of *V. pourtalesi*. The regression random forest model predicted zero to low (> 0 - 2.31 kg) biomass of this species across the much of the region. Predicted biomass was highest (up to 48.99 kg) in Emerald and LaHave Basins, coinciding with the location of the highest mean biomass values from the multispecies trawl surveys. Unlike the classification random forest model on presence-absence data, there was little to no extrapolation of predicted biomass beyond this area, and biomass was predicted to be low (> 0 to 4.03 kg) even in areas where medium-range catches (up to 15.10 kg) occurred.

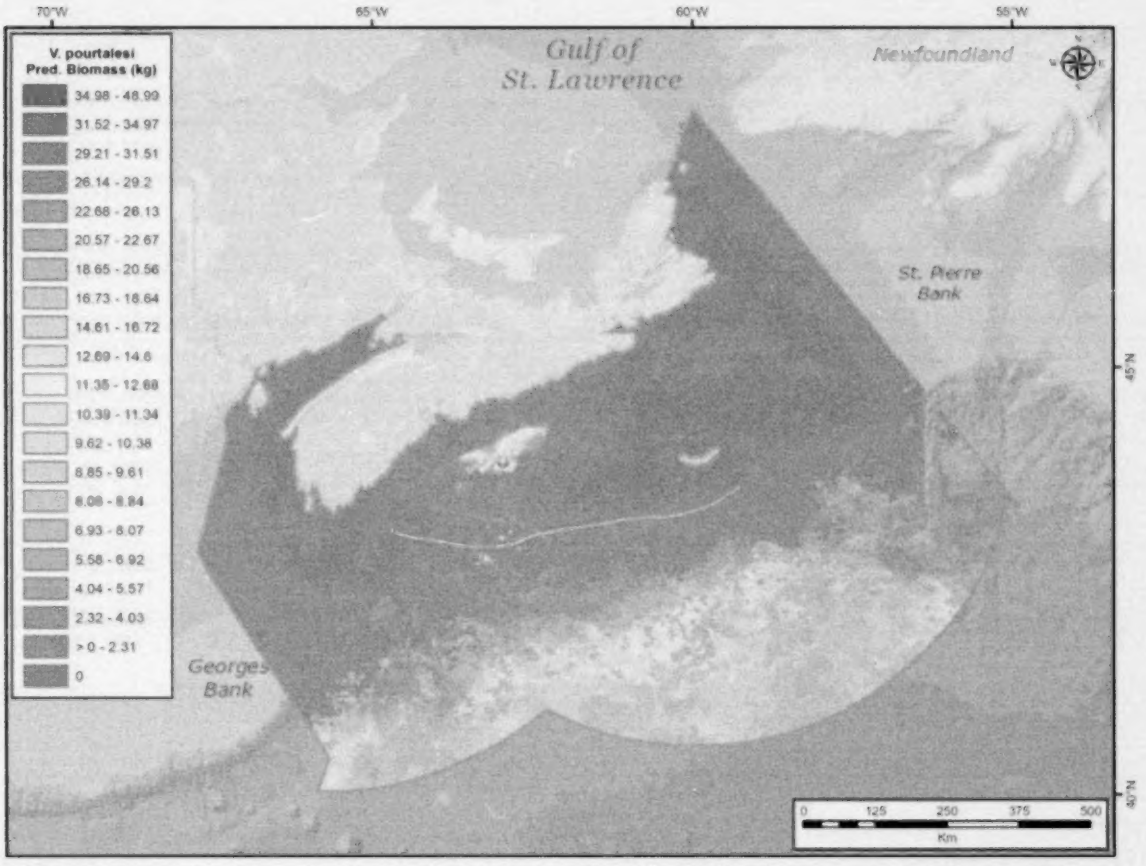

Figure 41. Predictions of biomass (kg) per grid cell of *Vazella pourtalesi* from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2007 and 2015.

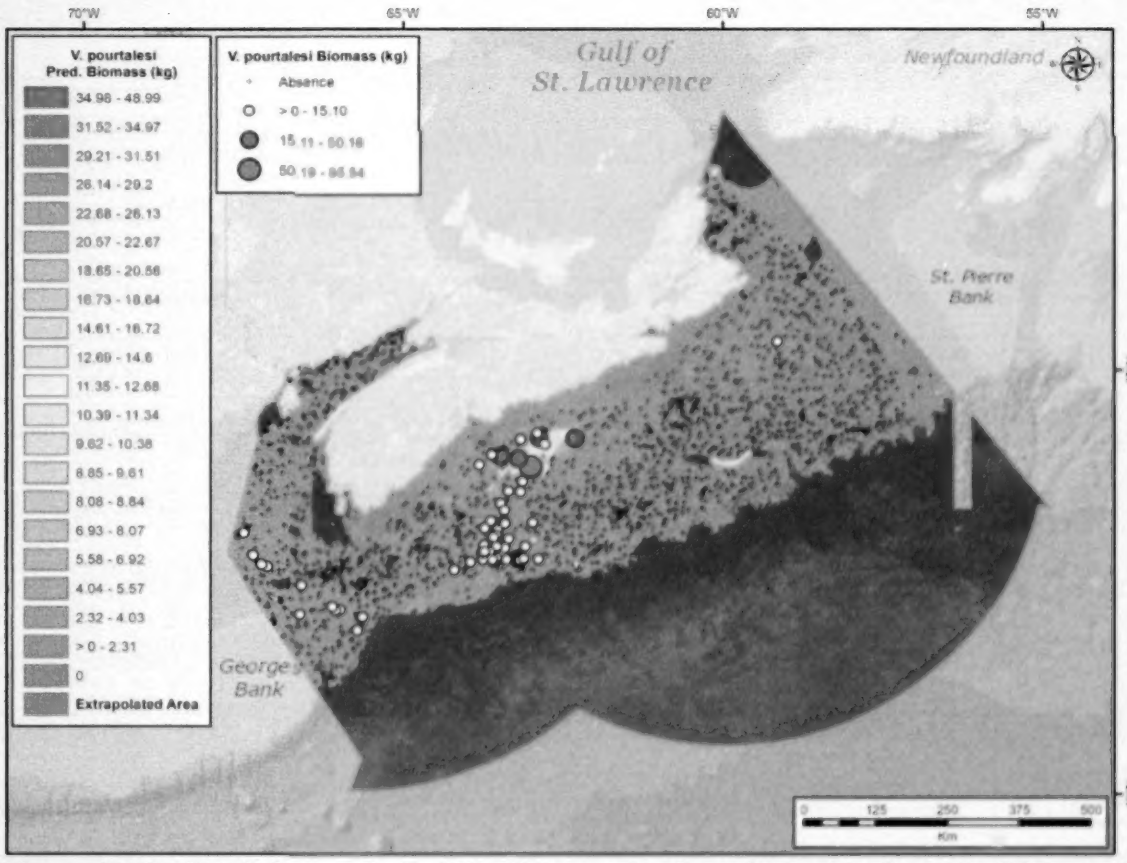

Figure 42. Predicted biomass (kg) of *Vazella pourtalesi* from the random forest model on catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2007 and 2015. Also shown are the mean biomass records per grid cell of *V. pourtalesi* from DFO multispecies trawl surveys and areas of model extrapolation.

Figure 43 shows the top 15 most important variables for predicting the *V. pourtalesi* biomass data. Bottom Temperature Average Minimum was the most important variable in this model. Prior to spatial interpolation, this variable displayed a slightly bimodal distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed a spatial pattern to data points over- and under-predicted by a normal distribution, with over-predicted points located in the Gulf of Maine, Emerald Basin, and on Banquereau and Misaine Banks, and in the deepest regions of the study extent, and under-predicted points located in Bay of Fundy, off southwestern Nova Scotia, the Laurentian Channel, and just beyond the shelf break. Bottom Temperature Average Minimum was followed very distantly in terms of its Mean Decrease in Gini Value by Bottom Salinity Average Maximum and Bottom Salinity Mean. Partial dependence of *V. pourtalesi* biomass on the top 6 environmental variables is shown in Figure 44. Predicted

biomass was highest at Bottom Temperature Average Minimum values greater than 8°C. Values in this range coincided with those data points over- and under-predicted by a normal distribution in Emerald Basin and in the Northeast Channel. The fit between predicted and observed values was poor, with severe under-prediction of temperature values greater than 8°C. Some points could therefore be predicted lower than their true values and slightly outside the range of highest predicted biomass identified in the partial plot.

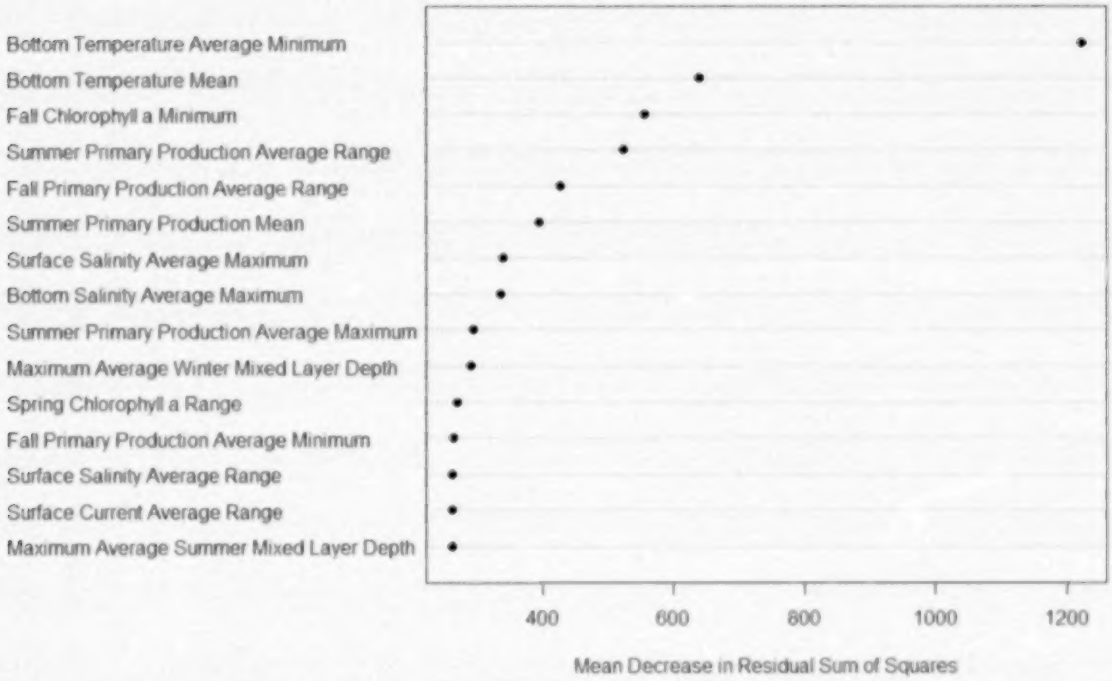

Figure 43. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on *Vazella pourtalesi* biomass per grid cell. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

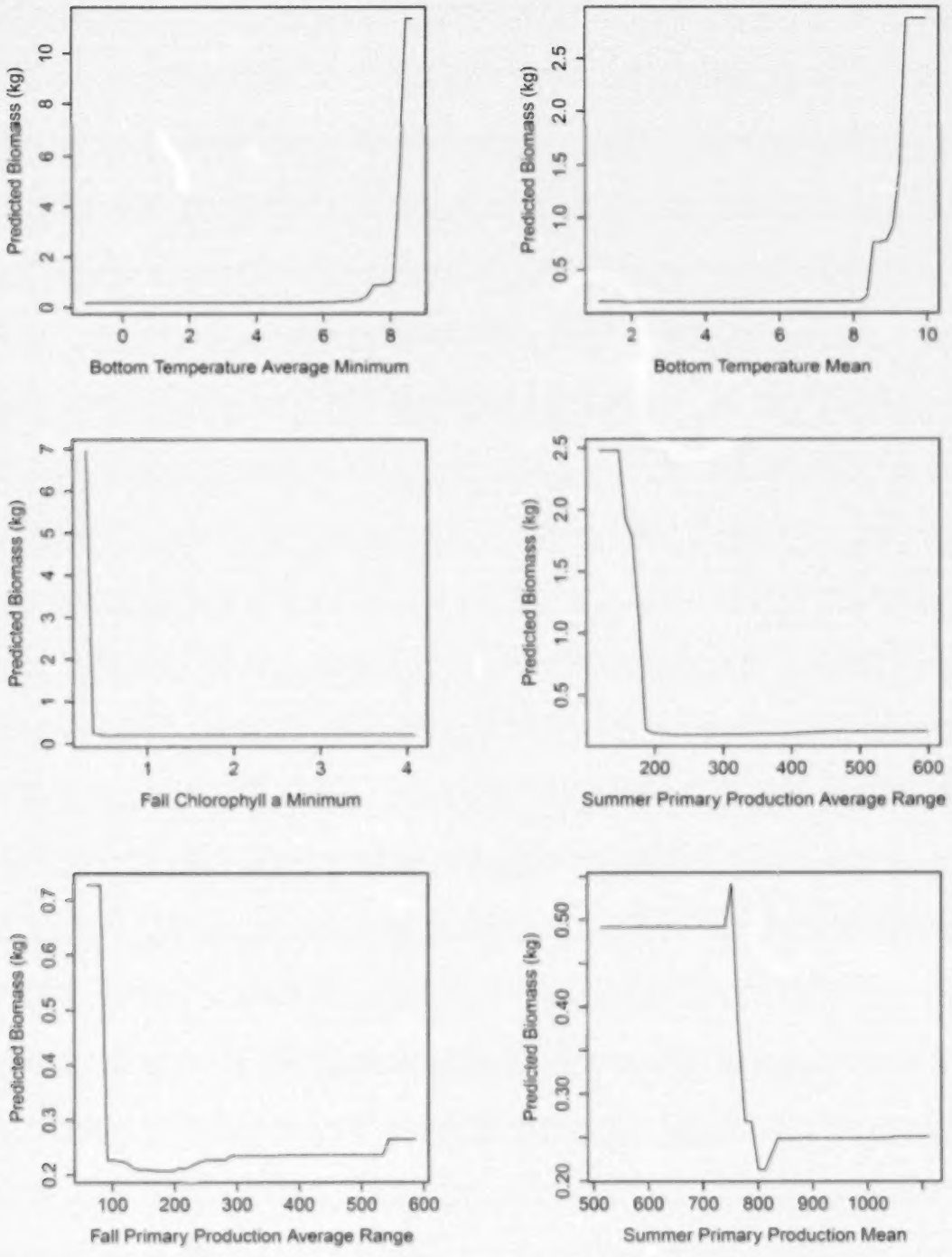

Figure 44. Partial dependence plots of the top six predictors from the random forest model of *Vazella pourtalesi* biomass data collected within the Maritimes Region, ordered left to right from the top. Predicted biomass (kg) is shown on the *y*-axis.

Sea Pens (Pennatulacea)

Data Sources and Distribution

Figure 45 shows the distribution of available sea pen records in the Maritimes Region. There was little overlap in the spatial distribution of records originating from the different data sources. DFO multispecies trawl survey records were concentrated in the central and eastern portions of the study area. Relatively few multispecies trawl records occurred along the slopes where the scientific survey and NOAA data were concentrated.

Initial random forest models of sea pens were run using only catch data originating from DFO multispecies trawl surveys (Western IIA gear). This data was collected over a period of

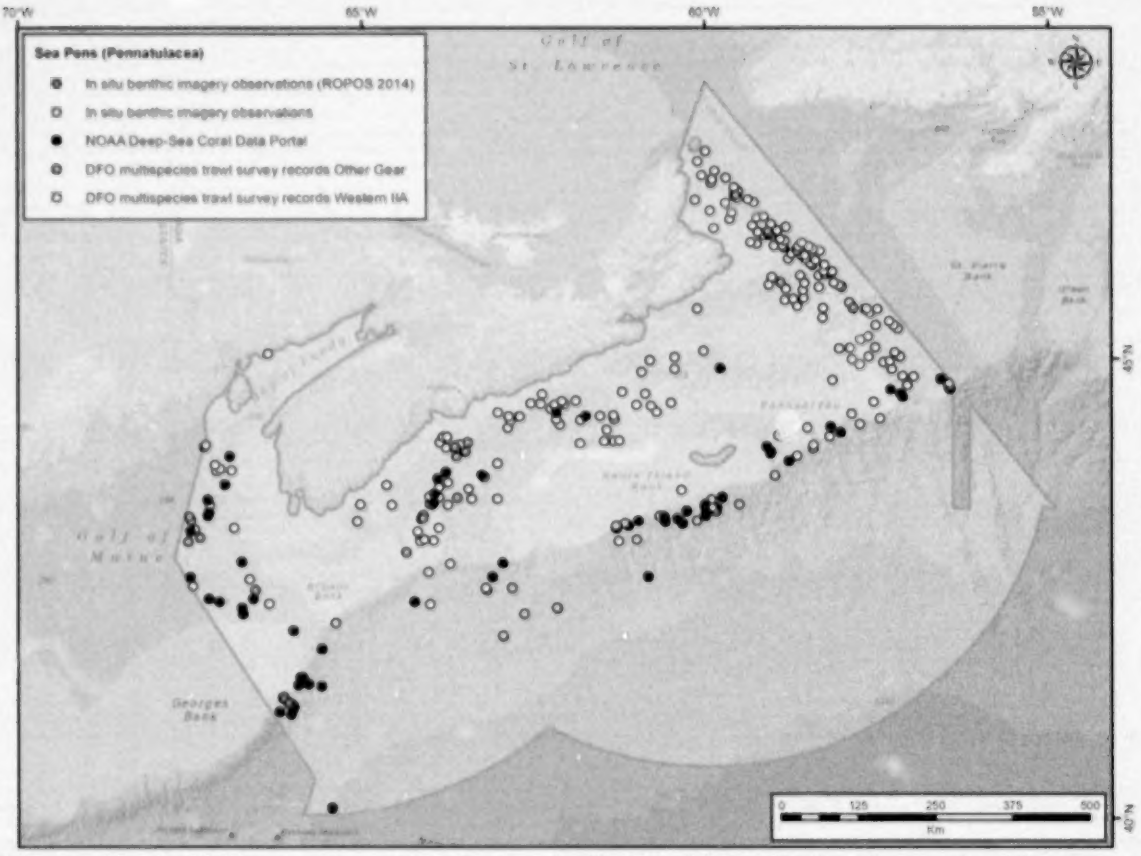

Figure 45. Available sea pen presence data in the Maritimes Region from scientific survey missions, the NOAA Deep-Sea Coral Data Portal, and DFO multispecies trawl surveys.

13 years from 2002 to 2014 (Table 13). This dataset consisted of 199 presence and 2710 absence records (Figure 46). Absence records were distributed relatively evenly across the study area. The highest mean biomass records (up to 30.62 kg) occurred in the Laurentian Channel. Smaller catches were distributed on central Scotian Shelf and in the Gulf of Maine.

Table 13. Number of presence and absence records of sea pen catch recorded from DFO multispecies trawl surveys conducted between 2002 and 2014 in the Maritimes Region.

Year	Total number of presences	Total number of absences
2002	2	110
2003	3	213
2004	1	80
2005	7	101
2006	10	282
2007	16	242
2008	15	270
2009	22	175
2010	36	263
2011	25	230
2012	24	239
2013	24	273
2014	14	232

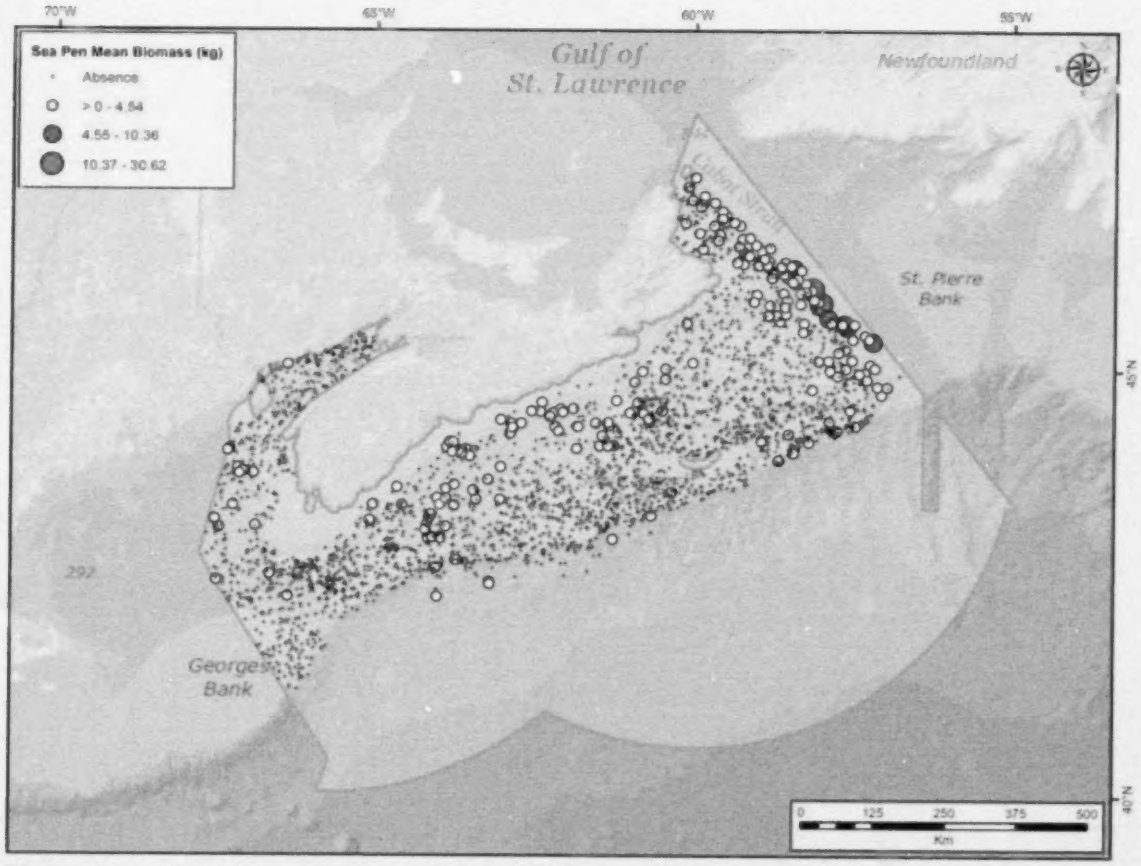

Figure 46. Mean biomass (kg) per grid cell of sea pen catch recorded from DFO multispecies trawl surveys from 2002 to 2014 within the Maritimes Region. Also shown are absence records from the same surveys.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity, and specificity) for the random forest model on balanced species prevalence (199 presences and 199 absences; Model 1) are presented in Table 14. The highest mean AUC of 0.894 was associated with Model Run 5 and is therefore considered the optimal model for the prediction of the sea pen response data. The sensitivity and specificity measures of this model run were 0.774 and 0.794, respectively. The confusion matrix of the optimal model is also presented in Table 2. Class error for both the presence and absence classes was somewhat moderate (0.226 and 0.206, respectively).

Table 14. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of sea pens within the Maritimes Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 5) which is considered the optimal model for predicting the presence probability of sea pens in the region.

Model Run	AUC	Sensitivity	Specificity
1	0.868	0.784	0.799
2	0.825	0.719	0.779
3	0.857	0.749	0.794
4	0.871	0.779	0.814
5	**0.894**	**0.774**	**0.794**
6	0.874	0.754	0.794
7	0.819	0.683	0.714
8	0.889	0.774	0.774
9	0.853	0.744	0.779
10	0.827	0.729	0.789
Mean	**0.858**	**0.749**	**0.783**
SD	**0.026**	**0.032**	**0.027**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	158	41	199	0.206
Presence	45	154	199	0.226

The presence probability prediction surface of sea pens is presented in Figure 47. The highest predictions of presence probability occurred on eastern Scotian Shelf and in Laurentian Channel. LaHave and Emerald Basins also had a high predicted presence probability of sea pens. Moderate to high sea pen presence probability occurred along the slopes and in eastern Gulf of Maine. These areas corresponded well with the spatial distribution of presence records (see Figure 48). Most of Bay of Fundy was predicted to have zero or low presence probability of sea pens.

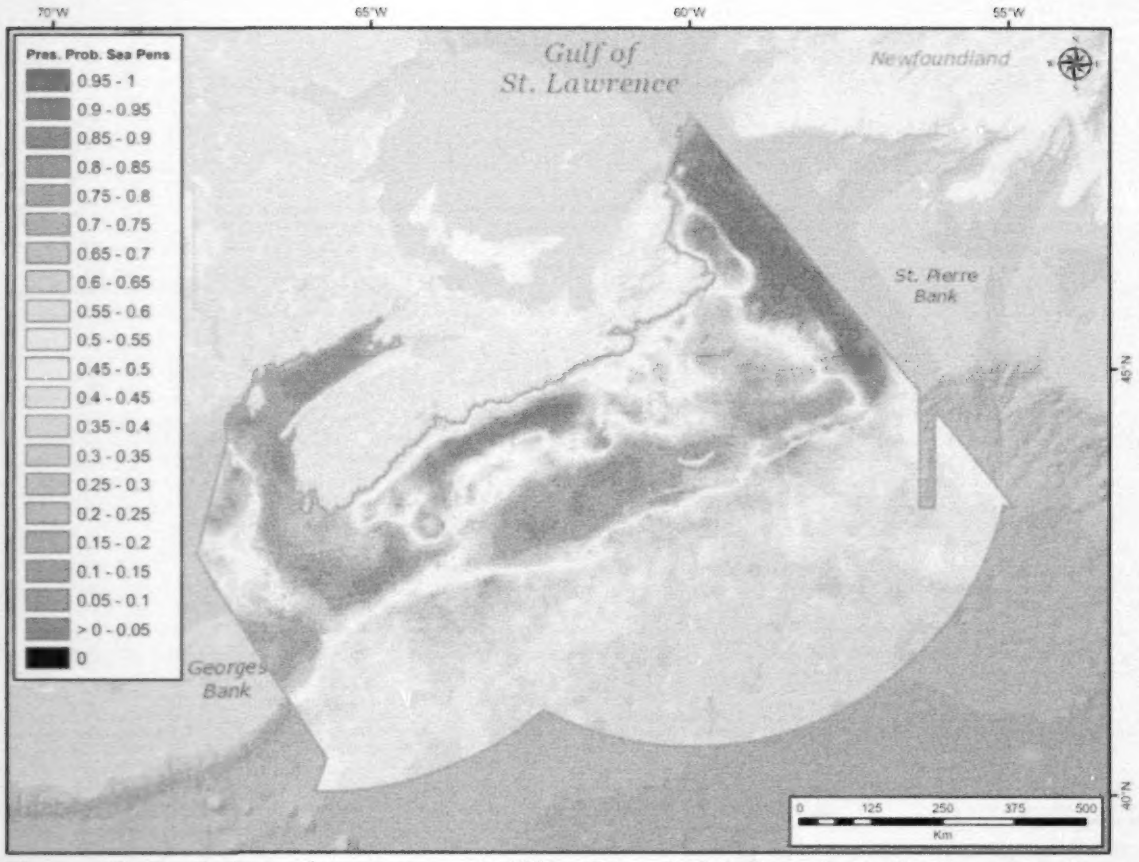

Figure 47. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of sea pen presence and absence data collected from DFO multispecies trawl surveys in the Maritimes Region between 2002 and 2014.

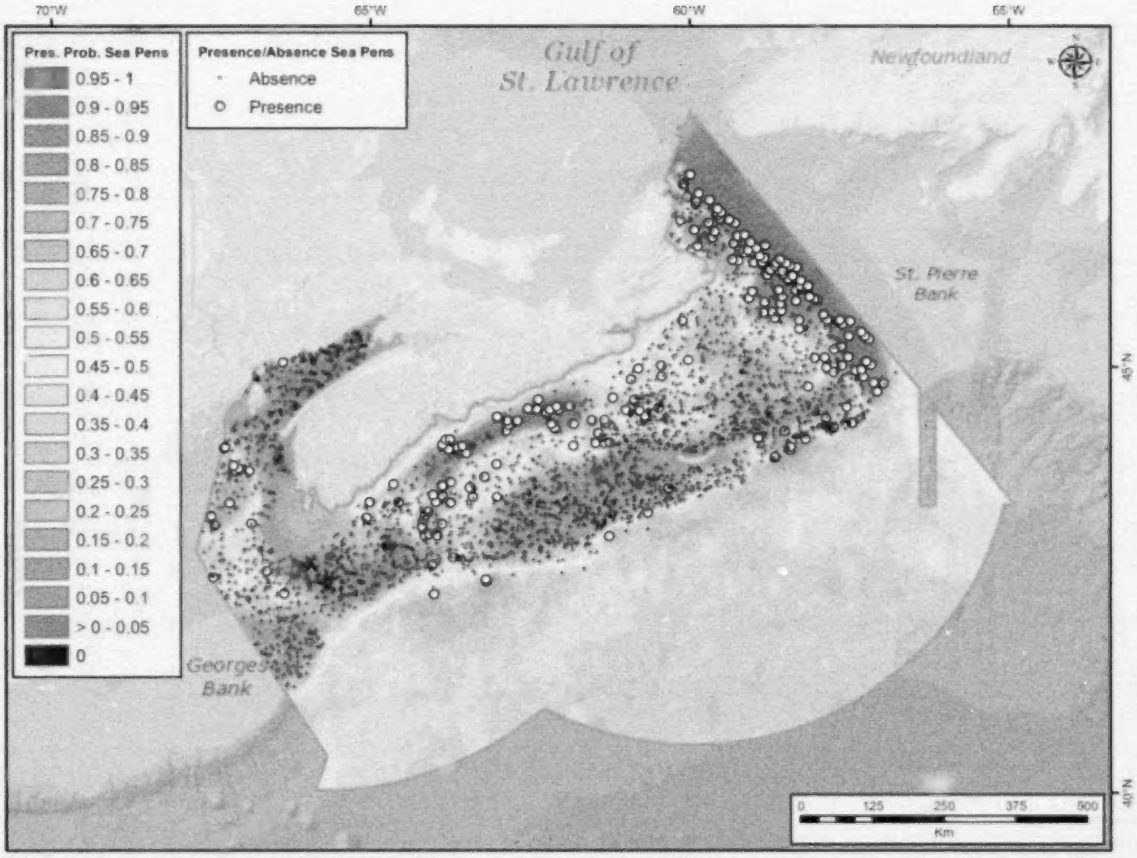

Figure 48. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of sea pen presence and absence data recorded from DFO multispecies trawl surveys in the Maritimes Region between 2002 and 2014.

The actual presence and absence data observations used in the optimal model run of Model 1 (199 presences and 199 absences; Figure 49) showed some spatial bias across the study area. Despite their being absence records in the eastern Gulf of Maine and in LaHave and Emerald Basins, very few were selected from these areas during the random down-sampling of the data prior to modelling. This likely caused an over-extension of high predicted probabilities in these areas. Areas of model extrapolation (i.e. areas where at least one environmental variable has values beyond its sampled range) are also shown in Figure 49. All deep water beyond the Scotian Shelf is considered extrapolated area, as well as smaller areas off southwestern Nova Scotia, central Scotian Shelf, and the northeastern tip of Cape Breton. The large extrapolated area off the northeast tip of Cape Breton has a high predicted presence probability of sea pens.

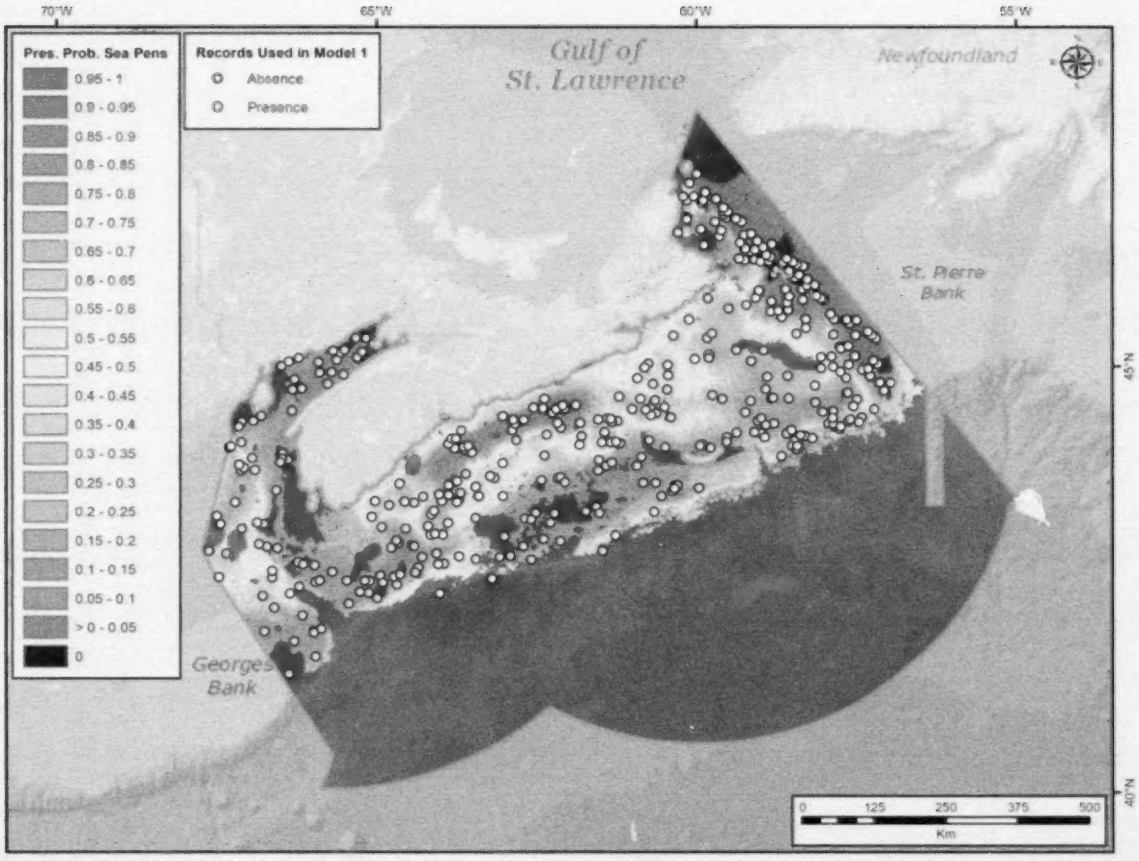

Figure 49. The 398 data observations (199 presences and 199 absences) of sea pens used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of sea pens generated frc.n Model 1.

Of all 66 environmental predictor variables used in the model, Surface Temperature Average Minimum was the most important for the classification of the sea pen presence and absence data (Figure 50). Prior to spatial interpolation, this variable displayed a right-skewed distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed a very strong spatial pattern to data points over- and under-predicted by a normal distribution, with over-predicted points located along the central and eastern coast of Nova Scotia and in the deepest regions of the study extent, and under-predicted points located in the centre of the study extent and in Bay of Fundy and Gulf of Maine. Surface Temperature Average Minimum was followed in terms of its Mean Decrease in Gini Value by Bottom Temperature Average Range and Depth. Surface and bottom current and salinity variables ranked high in the model. Partial dependence plots for the top 6 predictor variables are shown in Figure 51. The highest predicted sea pen presence

probabilities are associated with Surface Temperature Average Minimum values between -1 and 1°C. Values in this range coincided with those over-predicted values along the central and eastern coast of Nova Scotia and in Laurentian Channel. These are not of concern however, as there was a near-perfect fit between predicted and observed values in the interpolation with only slight over-prediction of temperature values between -1 and 1°C. A steep decrease in sea pen presence probability occurred at 4°C along the Bottom Temperature Average Range variable. Along the Depth gradient, sea pen presence probability sharply increased at ~200 m depth.

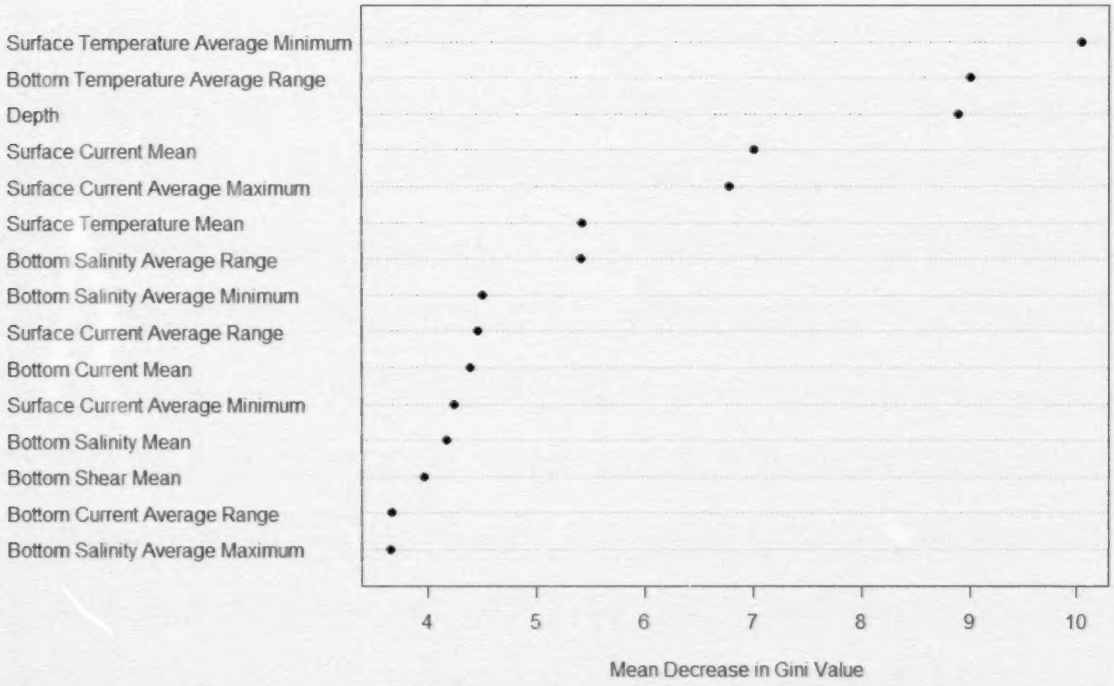

Figure 50. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting sea pen presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

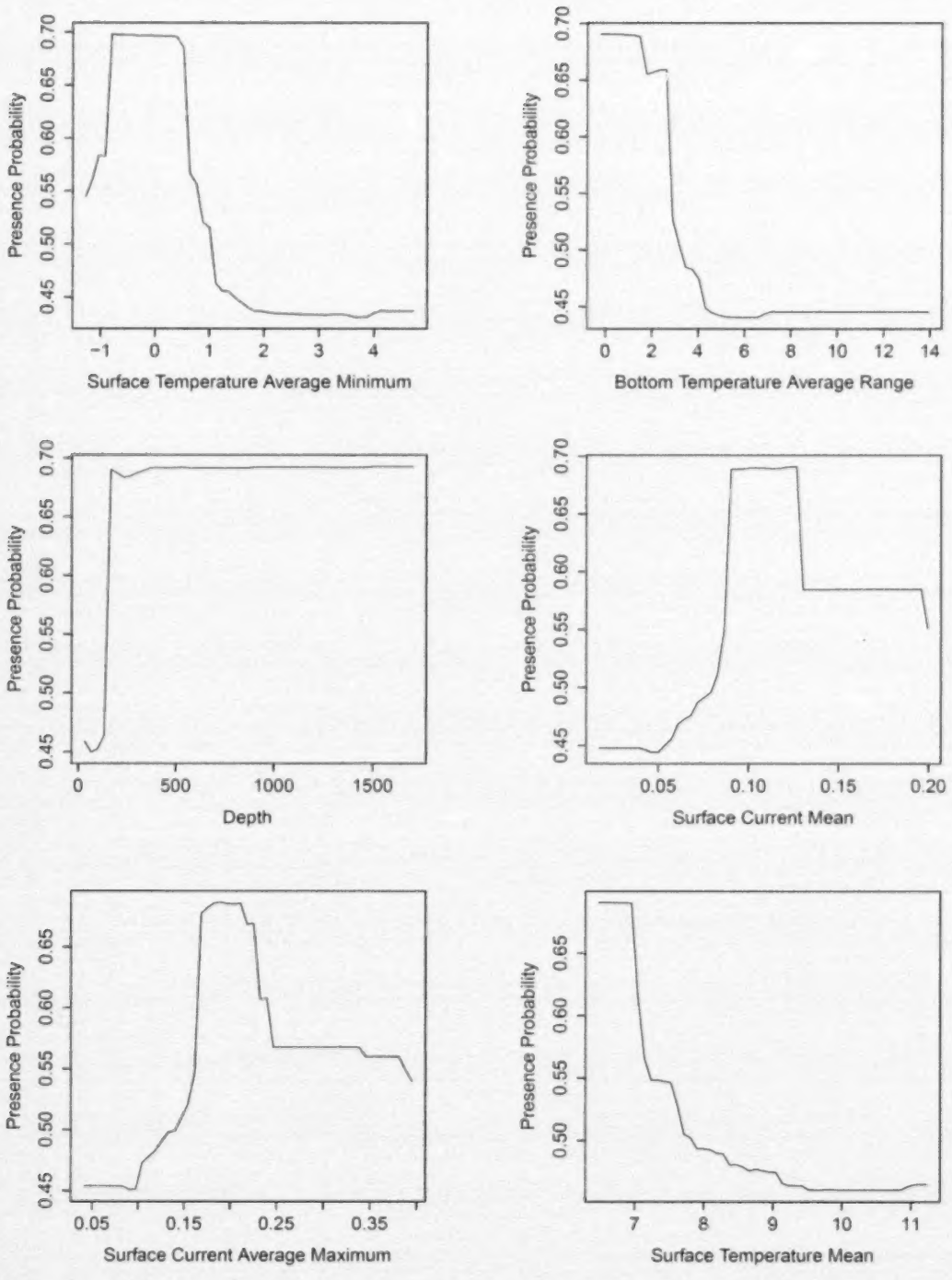
Presence Probability
0.45 0.50 0.55 0.60 0.65 0.70
−1 0 1 2 3 4
Surface Temperature Average Minimum
Presence Probability
0.45 0.50 0.55 0.60 0.65 0.70
0 2 4 6 8 10 12 14
Bottom Temperature Average Range
Presence Probability
0.45 0.50 0.55 0.60 0.65 0.70
0 500 1000 1500
Depth
Presence Probability
0.45 0.50 0.55 0.60 0.65 0.70
0.05 0.10 0.15 0.20
Surface Current Mean
Presence Probability
0.45 0.50 0.55 0.60 0.65
0.05 0.15 0.25 0.35
Surface Current Average Maximum
Presence Probability
0.50 0.55 0.60 0.65
7 8 9 10 11
Surface Temperature Mean

Figure 51. Partial dependence plots of the top six predictors from the optimal random forest model of sea pen presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 15 shows the accuracy measures for the random forest model on all sea pen presence and absence data (199 presences and 2710 absences) and a threshold equal to species prevalence (0.07). The average AUC calculated from 10-fold cross validation was 0.857, nearly identical to the average AUC from Model 1. Class error for the absence class was similar to that of Model 1 (0.200 compared to 0.206 from Model 1), however class error for the presence class was slightly higher (0.261 compared to 0.226 from Model 1). Sensitivity and specificity measures were comparable to those of Model 1.

The surface of predicted presence probability of sea pens generated from Model 2 is much more conservative than that of Model 1 (Figure 52). Similar to Model 1, the highest presence probability of sea pens in Model 2 occurred in the Laurentian Channel. The high presence probability St. Ann's Banks, and Misaine and Banquereau Banks in Model 1 were much reduced in this model. After the Laurentian Channel, Emerald Basin had the highest predicted probability of presence of sea pens. Much of the Bay of Fundy, Browns Bank, and Sable Island Bank were predicted to have zero presence of sea pens. On the Scotian Shelf, areas of higher sea pen presence probability did not extend far beyond the location of the presence records (see Figure 53).

Table 15. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of sea pens within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.879		**Absence**	**Presence**				
2	0.946	**Absence**	2171	539	2710	0.200	0.739	0.801
3	0.774	**Presence**	52	147	199	0.261		
4	0.884							
5	0.828							
6	0.896							
7	0.853							
8	0.899							
9	0.796							
10	0.821							
Mean	**0.857**							
SD	**0.053**							

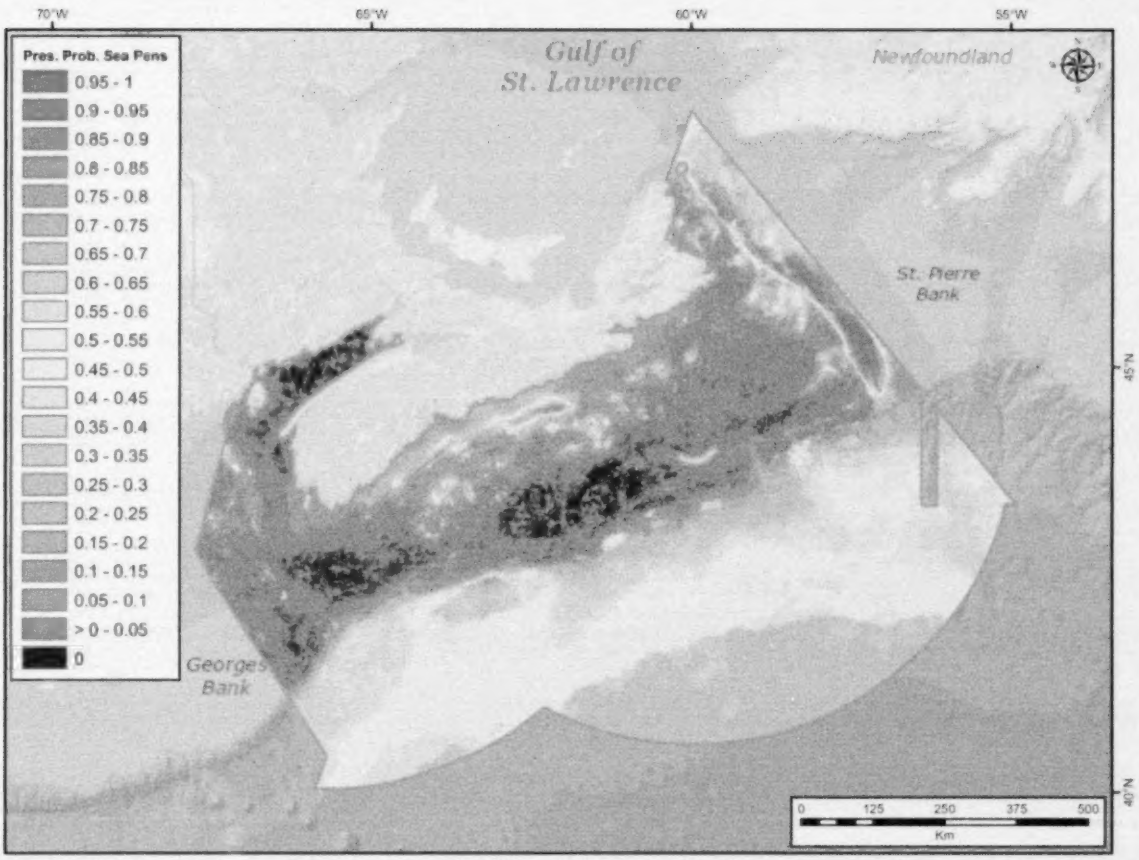

Figure 52. Predictions of presence probability (Pres. Prob.) of sea pens based on a random forest model on unbalanced presence and absence sea pen catch data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2002 and 2014.

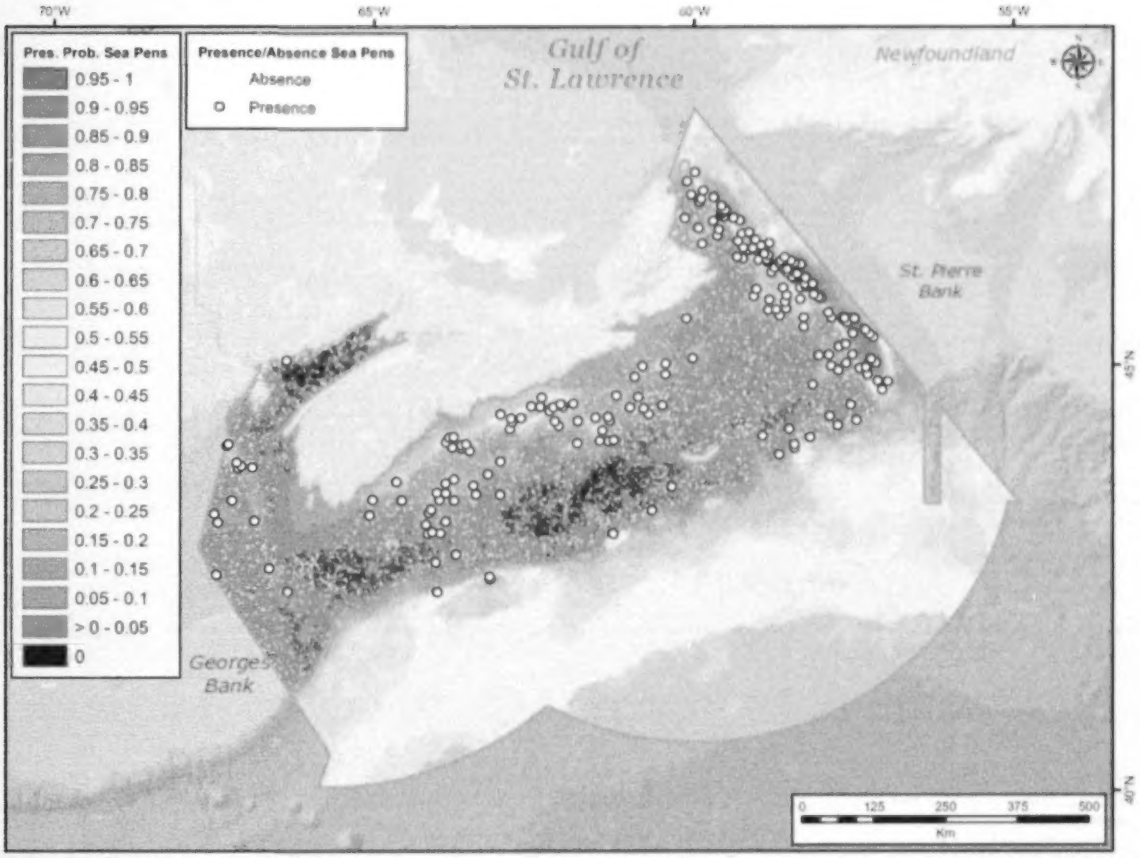

Figure 53. Presence and absence observations and predictions of presence probability (Pres. Prob.) of sea pens based on a random forest model on unbalanced presence and absence sea pen catch data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2002 and 2014.

Areas of extrapolation of the unbalanced random forest model on sea pens is presented in Figure 54. Areas of extrapolation, particularly the large area off the northeast tip of Cape Breton, were not associated with high predicted presence probabilities.

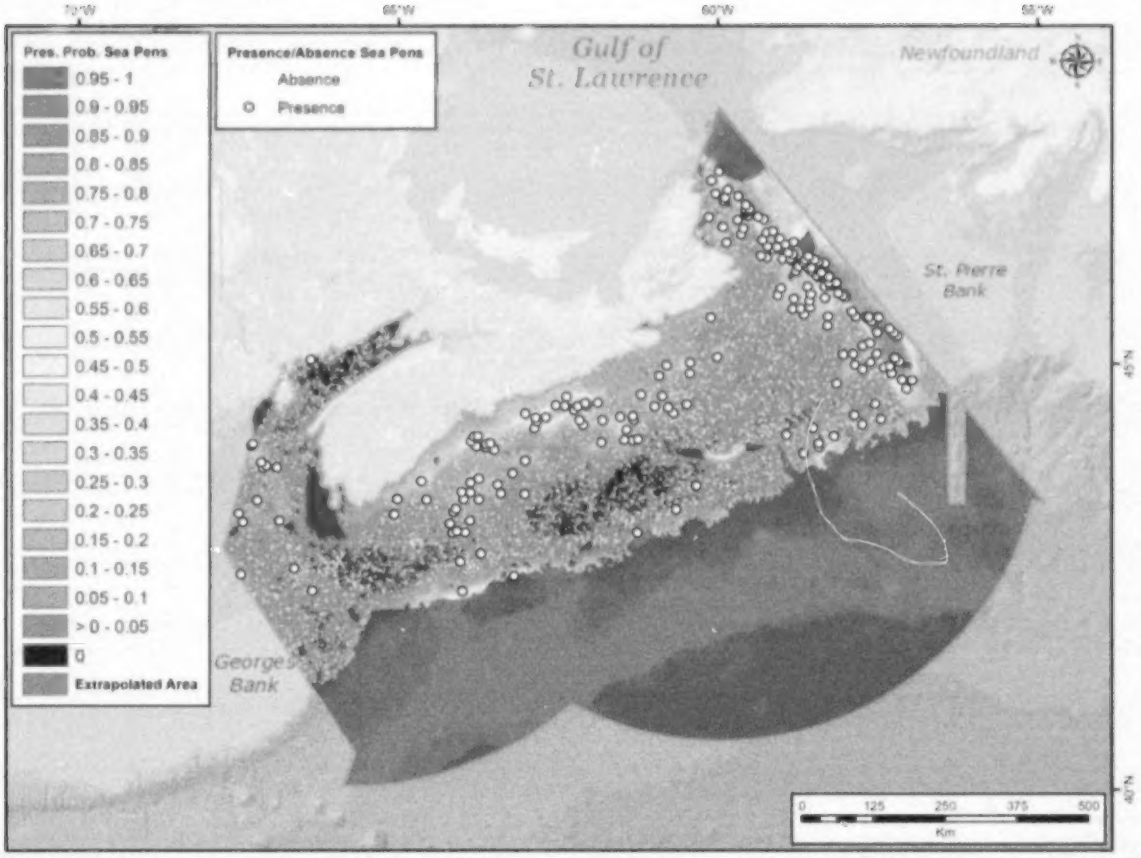

Figure 54. Areas of extrapolation of the random forest model on unbalanced presence and absence sea pen catch data collected within the Maritimes Region between 2002 and 2014. Also shown are the sea pen presence and absence observations and predictions of presence probability (Pres. Prob.).

The order of importance of the environmental predictor variables in Model 2 is slightly different from that of Model 1 (Figure 55). Surface Temperature Mean was the most important variable in Model 2, compared to Surface Temperature Average Minimum in Model 1. Prior to spatial interpolation, the Surface Temperature Mean variable displayed a right-skewed distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed a strong spatial pattern to data points over- and under-predicted by a normal distribution, with over-predicted points located along the central and eastern coast of Nova Scotia and in the deep regions of the study extent, and under-predicted points located in the centre of the study extent and in Bay of Fundy. Temperature Average Minimum was followed closely in terms of its Mean Decrease in Gini Value by Depth. Similar to Model 1, surface current and temperature variables, and bottom

salinity variables had higher Gini values than the chlorophyll *a* and primary production variables in the model. Partial dependence plots of the top 6 environmental predictor variables are shown in Figure 56. The highest predicted sea pen presence probabilities were associated with the lowest (< 8°C) and highest (> 11°) Surface Temperature Mean values. Values in this range coincided with those over-predicted values along the central and eastern coast of Nova Scotia and in Laurentian Channel, and in the deeper regions of the study extent. These are not of concern however, as there was a near-perfect fit between predicted and observed values in the interpolation with only slight over-prediction of temperature values in these ranges. Along the Depth gradient, sea pen presence probability steadily increased beginning at ~200 m and continued to increase in a step-like fashion.

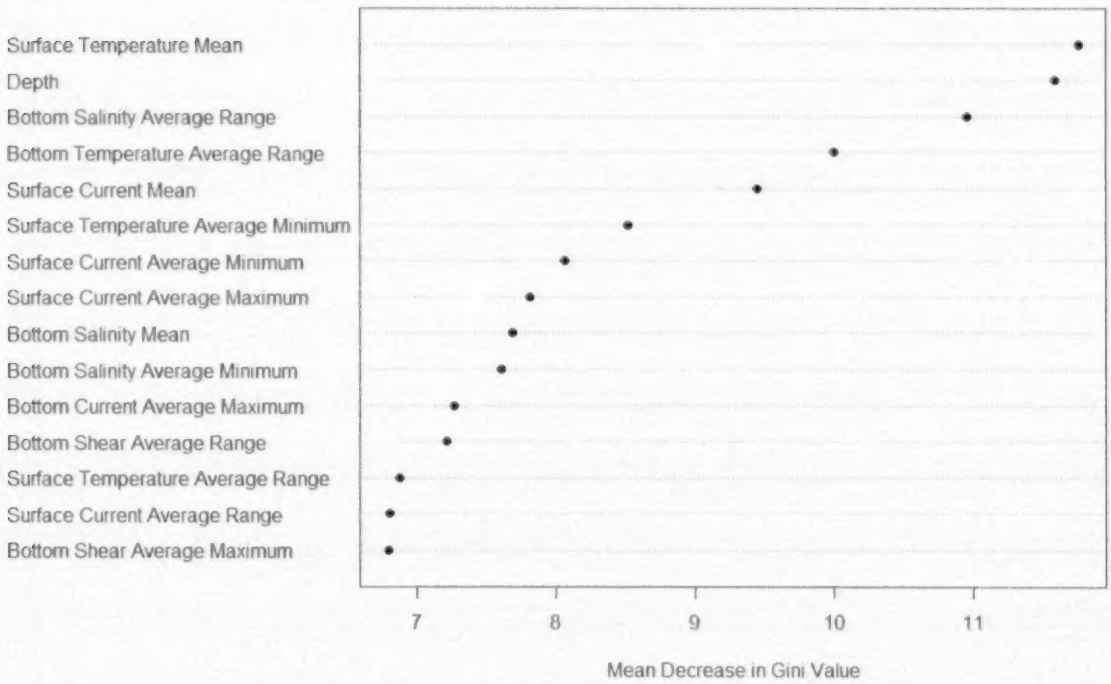

Figure 55. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced sea pen presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

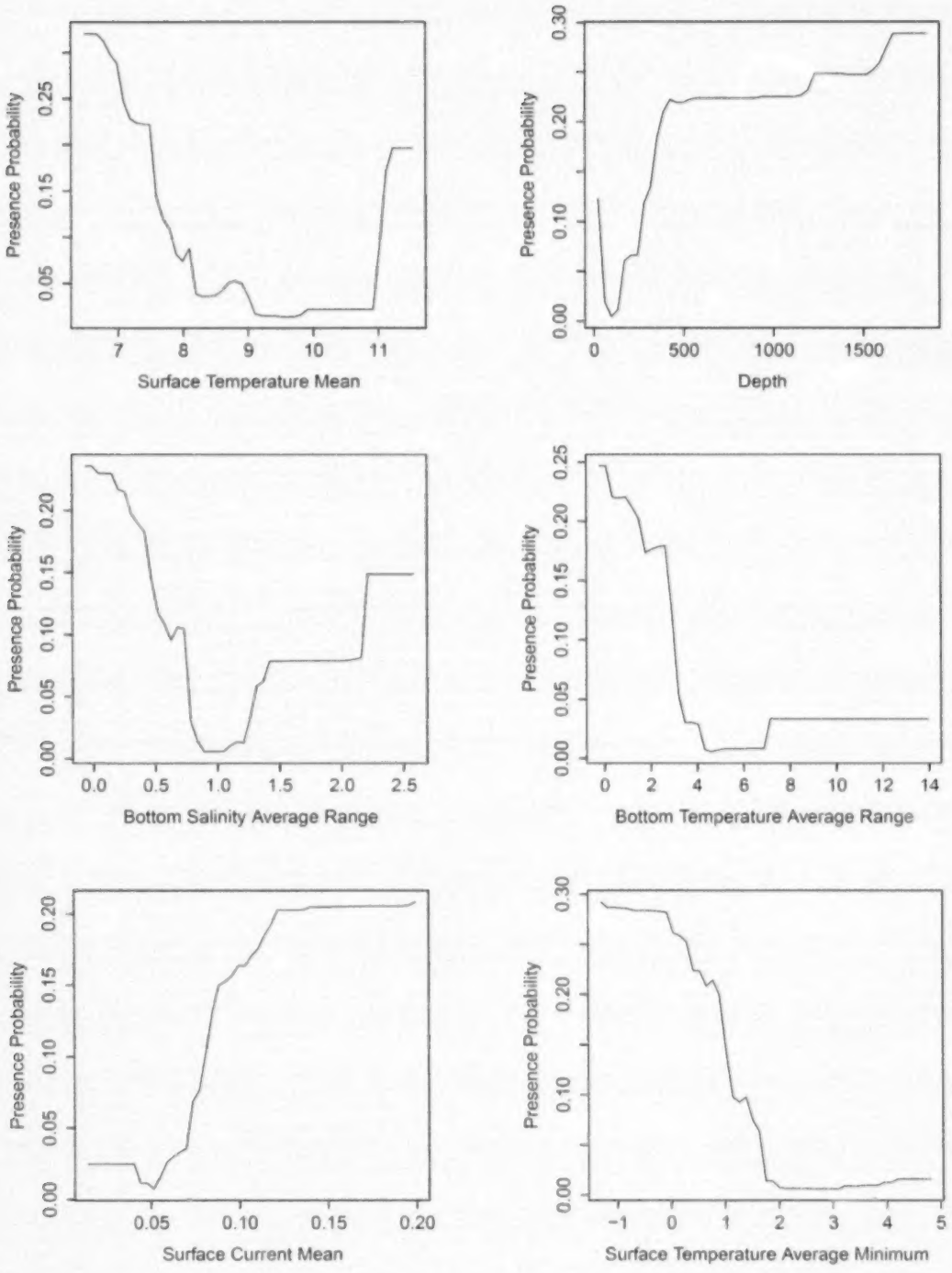

Figure 56. Partial dependence plots of the top six predictors from the random forest model of sea pen unbalanced presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 3 -- Addition of *In Situ* Benthic Imagery Observations

Given the relatively low number of sea pen records in the Maritimes Region, the DFO multispecies trawl survey data were augmented with additional *in situ* benthic imagery records from scientific surveys conducted in the Maritimes Region between 1965 and 2011 (see Figure 45). These data were collected on scientific missions led by DFO, NRCan, or Dalhousie University. A total of 149 additional presence records (Table 16) were added to the dataset after filtering the data so that only one presence record occurred per environmental grid cell. The combined dataset consisting of 348 presences and 2708 absences was remodelled (termed Model 3) using an unbalanced design and a threshold equal to species prevalence (0.11). The accuracy measures for random forest Model 3 are shown in Table 17. The average AUC computed from 10-fold cross validation was 0.901 ± 0.031 SD, the highest of all three models. Class error for the presence and absence classes was the lowest of all three models, while sensitivity and specificity were high.

Table 16. Number of *in situ* benthic imagery observations of sea pens collected from various surveys conducted within the Maritimes Region 1965 and 2014 in the Maritimes Region.

Year	Gear	Total number of presences
1965	NRCan Drop Camera	2
1967	NRCan Drop Camera	1
2000	NRCan Drop Camera	7
2001	ROPOS	6
2005	Campod	10
2006	ROPOS	16
2006	DSIS	3
2007	ROPOS	37
2008	NRCan Drop Camera	5
2008	Campod	53
2011	Campod	5
2014	Towed Camera	4

Table 17. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of sea pens from DFO multispecies trawl survey records and *in situ* benthic imagery observations collected within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.910		**Absence**	**Presence**				
2	0.868	**Absence**	2219	489	2708	0.181	0.813	0.819
3	0.920	**Presence**	65	283	348	0.187		
4	0.914							
5	0.945							
6	0.873							
7	0.876							
8	0.940							
9	0.911							
10	0.857							
Mean	**0.901**							
SD	**0.031**							

The additional presence records expanded the area of high sea pen presence probability along the eastern slope and in the deep-water canyons (Figure 57). The Gully submarine canyon east of Sable Island, and the Northeast Channel on the western Scotian Shelf showed much higher presence probability of sea pens compared to Models 1 and 2. These areas of higher presence probability along the slope corresponded well with the location of the additional *in situ* imagery records added to the model (Figure 58). The area of extrapolation along the slope off eastern Scotian Shelf is reduced with the addition of science survey presence records there (Figure 59). Figure 60 depicts the predicted distribution of sea pens based on a prevalence threshold of 0.11. In this map, all presence probability values generated from Model 3 that were greater than 0.11 were classified as presence, while values less than 0.11 were classed as absence. The majority of the slope and deep-water channels was classified as presence of sea pens. A large area off southwest Cape Breton is also classified as presence.

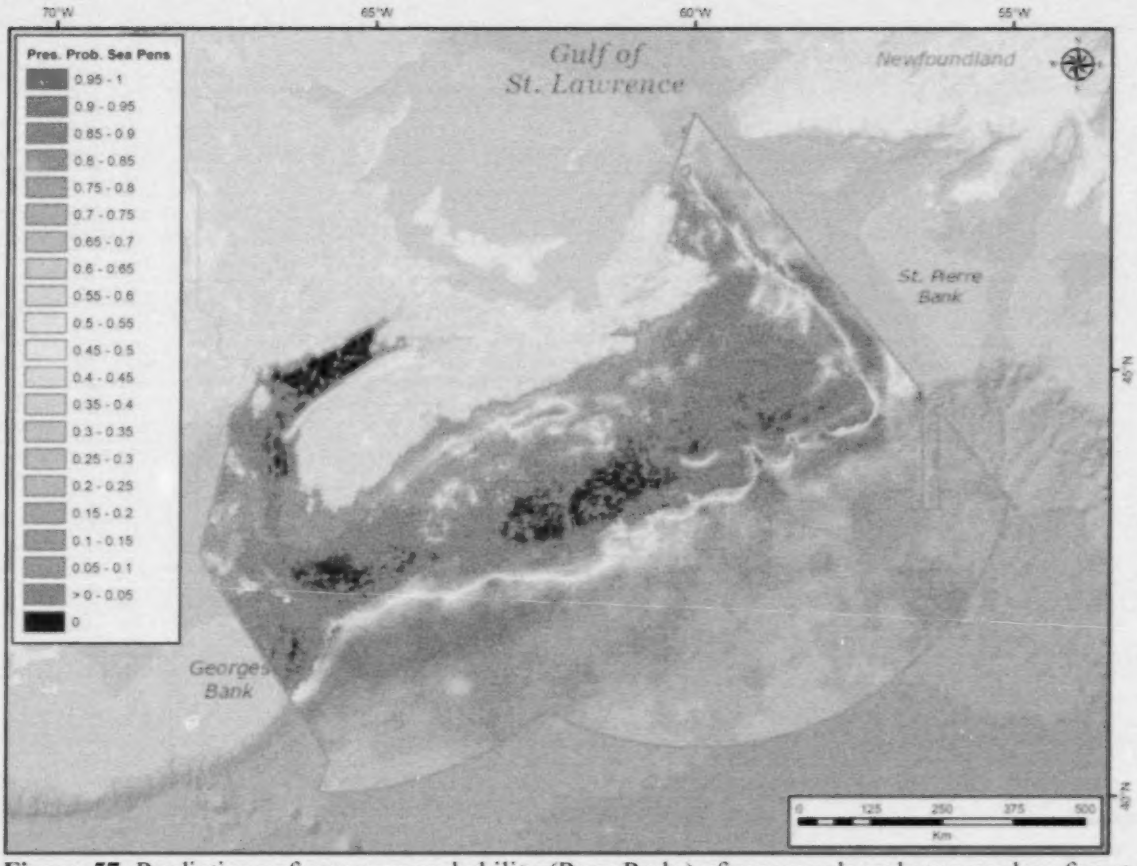

Figure 57. Predictions of presence probability (Pres. Prob.) of sea pens based on a random forest model on unbalanced presence and absence sea pen catch data from DFO multispecies trawl surveys and *in situ* benthic imagery observations of sea pens collected from various surveys conducted between 1965 and 2014 in the Maritimes Region.

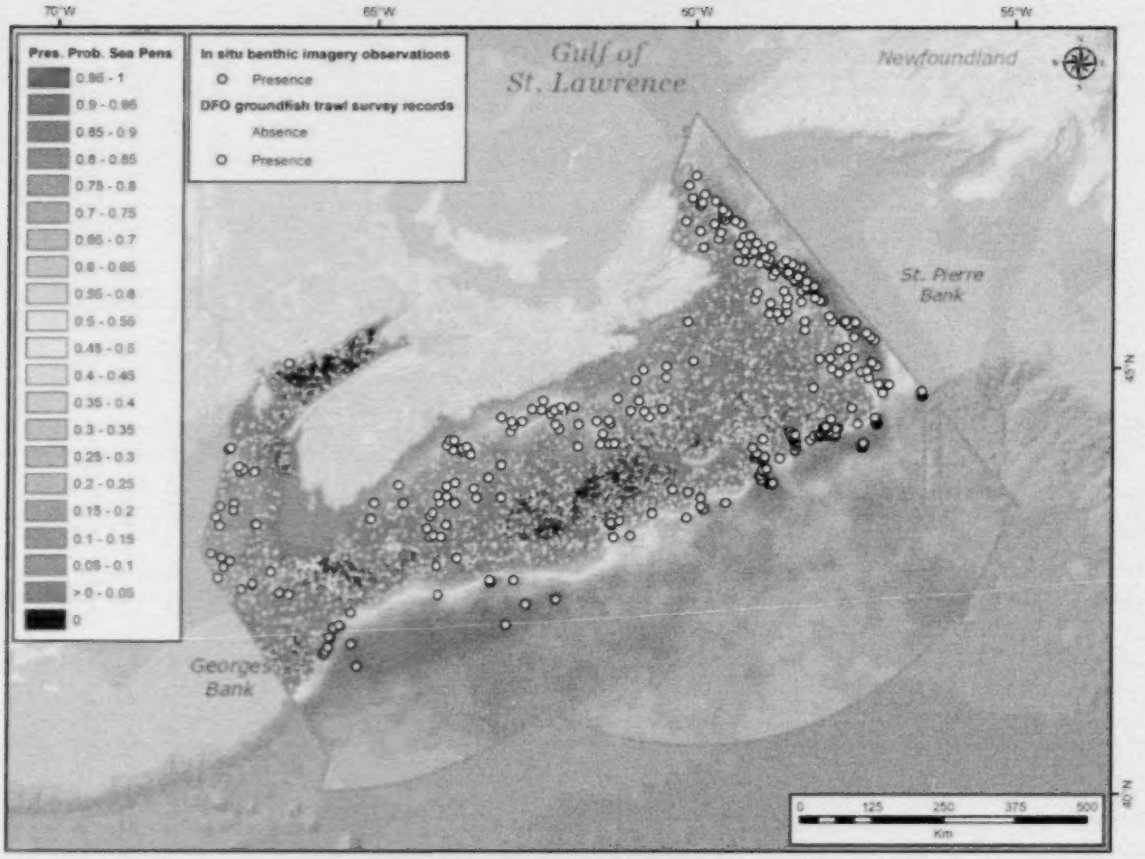

Figure 58. Presence and absence observations and predictions of presence probability (Pres. Prob.) of sea pens based on a random forest model on unbalanced presence and absence sea pen catch data from DFO multispecies trawl surveys and *in situ* benthic imagery observations of sea pens collected from various surveys conducted within the Maritimes Region 1965 and 2014.

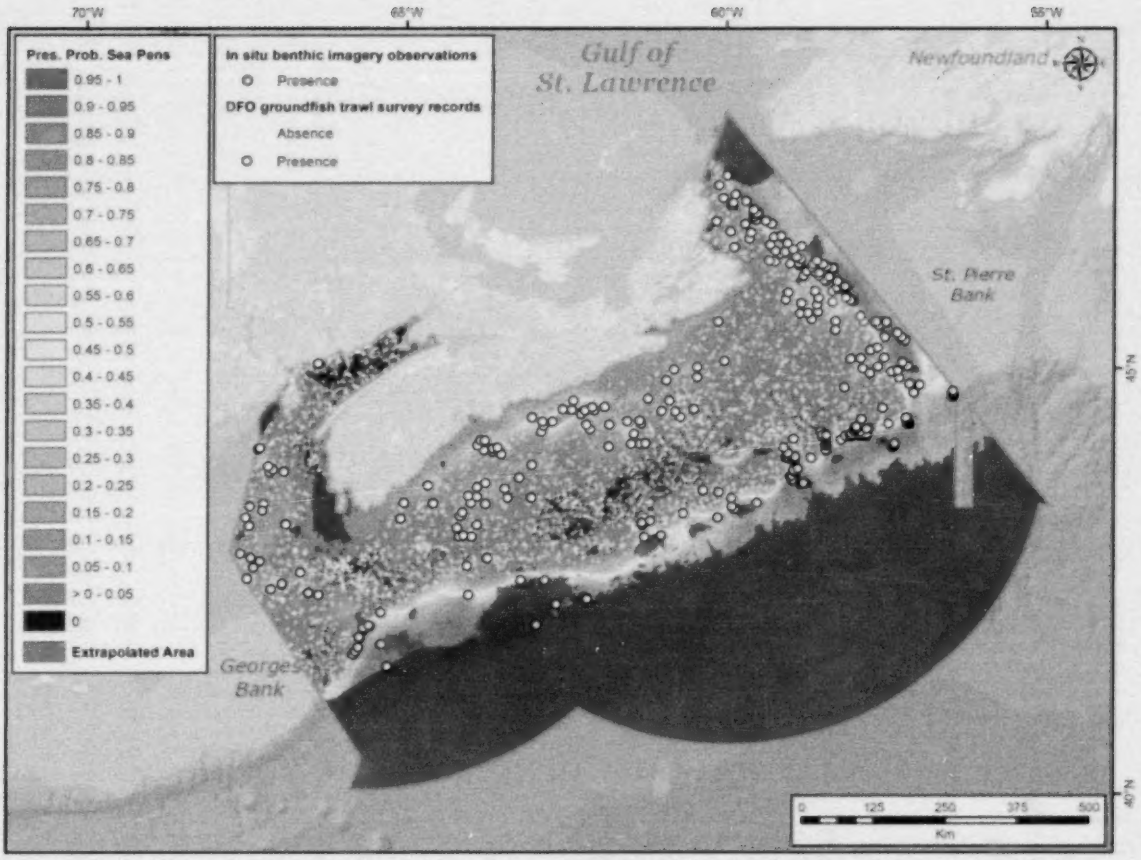

Figure 59. Areas of extrapolation of a random forest model on unbalanced presence and absence sea pen catch data from DFO multispecies trawl surveys and *in situ* benthic imagery observations of sea pens collected from various surveys conducted within the Maritimes Region 1965 and 2014. Also shown are the presence and absence observations and predictions of presence probability (Pres. Prob.).

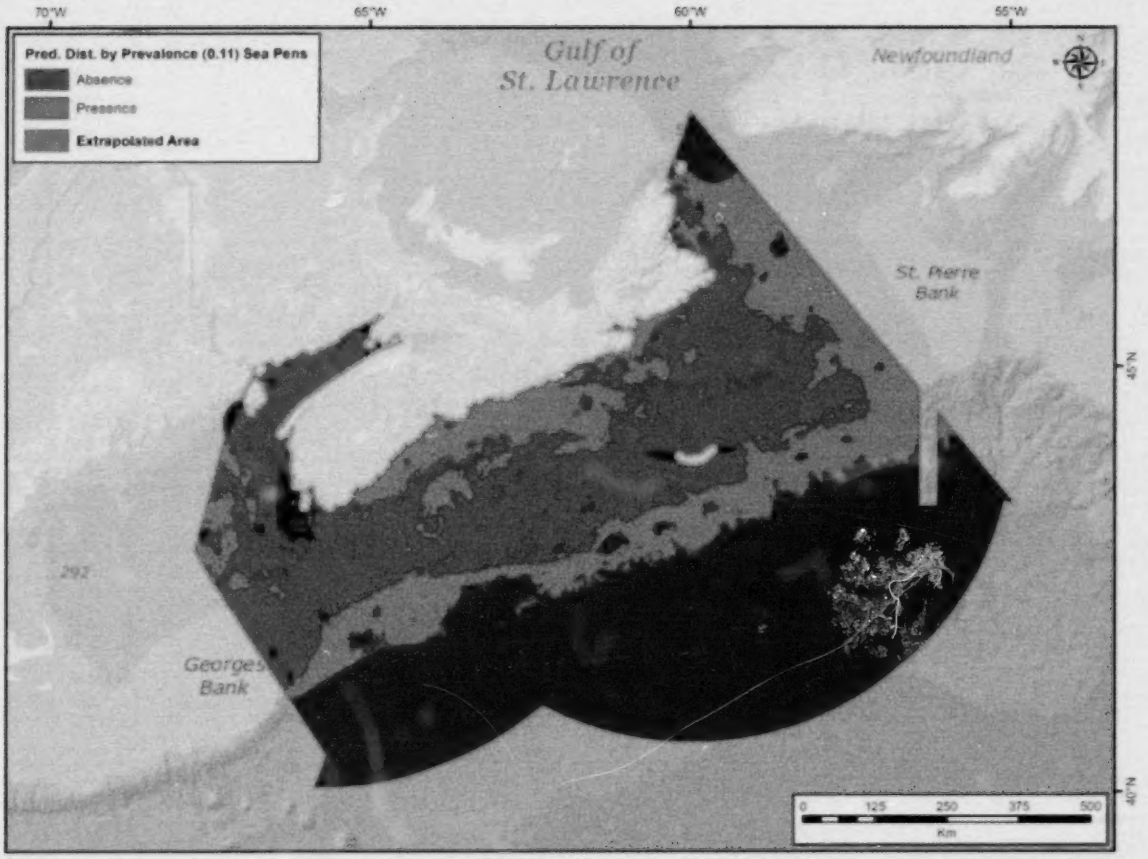

Figure 60. Predicted distribution (Pred. Dist.) of sea pens in the Maritimes Region based on the prevalence threshold of 0.11 of sea pen presence and absence data used in Model 3. Also shown are the areas of model extrapolation (grey polygon may appear red or blue).

In contrast to Model 1 and Model 2, Depth and Slope, two non-interpolated variables, were the top two predictors of sea pen presence in Model 3 (Figure 61). These variables were followed distantly by Bottom Temperature Average Range and the remaining variables in the model. Bottom salinity and surface current variables ranked high in this model. Partial dependence plots of the top 6 environmental variables are shown in Figure 62. Presence probability increased rapidly at the shallowest depths up to ~500 m, where it then plateaued. Probability of presence of sea pens increased rapidly along the Slope gradient and reached a plateau at ~10°.

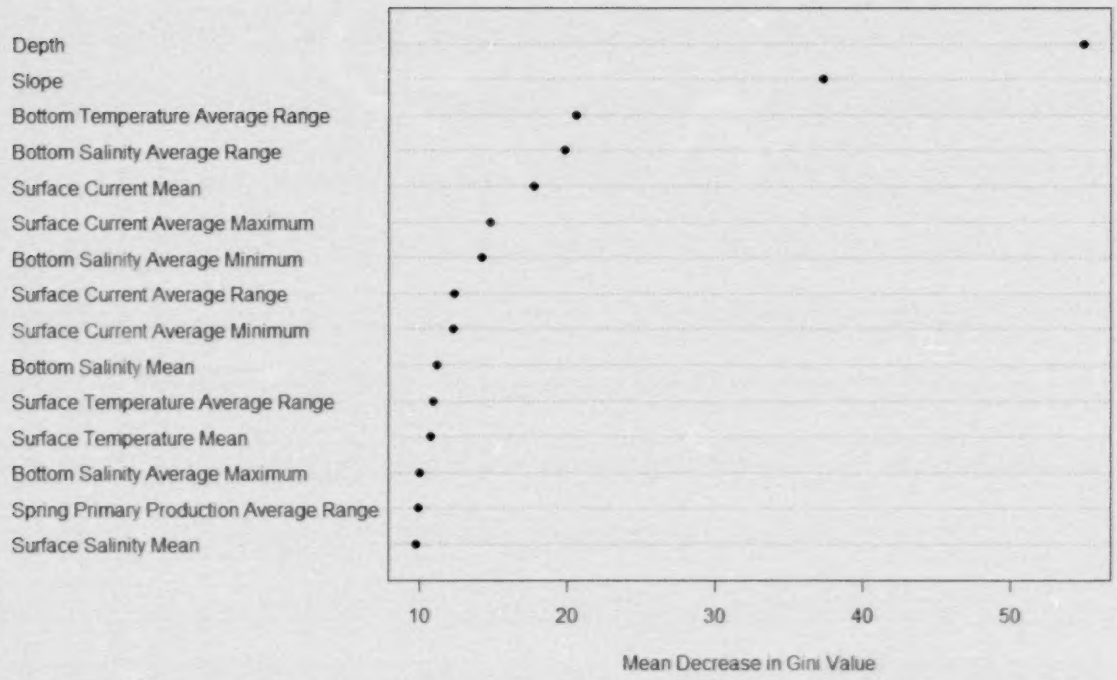

Figure 61. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced sea pen presence and absence data from DFO multispecies trawl surveys and *in situ* benthic imagery observations collected from various surveys conducted within the Maritimes Region between 1965 and 2014. The higher the Mean Gini value the more important the variable is for predicting the response data.

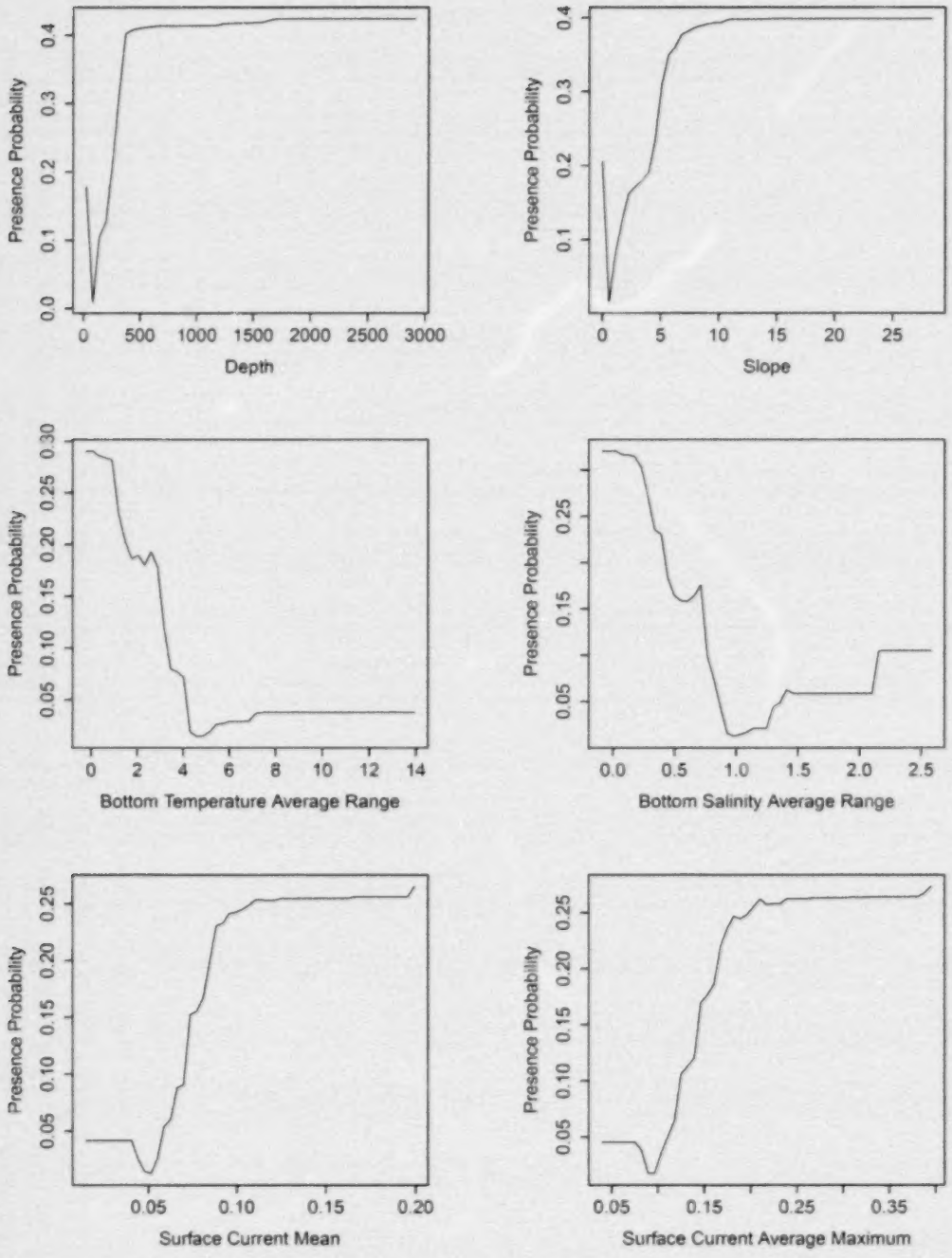

Figure 62. Partial dependence plots of the top six predictors from the random forest model of sea pen unbalanced presence and absence data from DFO multispecies trawl surveys and *in situ* benthic imagery observations collected from various surveys conducted within the Maritimes Region between 1965 and 2014, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all available sea pen records and an unbalanced species prevalence (Model 3) was selected as the best predictor of sea pen distribution in the Maritimes Region. Model 1 (balanced species prevalence) was considered a poor predictor of sea pen presence probability due to its exaggeration of high presence probability in Emerald and LaHave Basins, and in the Laurentian Channel where there were no presence records to support it. This phenomenon was likely due to random down-sampling of the absence data. Model 2, which was generated using the same presence-absence dataset but using all absence data, produced a much more realistic presence probability surface with less exaggeration beyond the location of presence points. The additional presence records added to Model 3 produced the highest AUC and sensitivity and specificity measures of all three models. Although the presence probability surface was similar to that of Model 2, this model predicted higher sea pen presence probability along the eastern Scotian Shelf slope and in its canyons, providing a more accurate depiction of the distribution of sea pens in the region based on the available data.

Validation of Selected Model Using Independent Data

Figure 63 shows the predicted presence probabilities of sea pens generated from Model 3 at the location of sea pen records from the NOAA Deep-Sea Coral Data Portal. Many of the NOAA records were concentrated in the various canyons along the Scotian Slope where the model predicted a high probability of occurrence of sea pens. Of the 98 sea pen records from this data source, 17% were predicted as absences based on the prevalence threshold of 0.11 (yellow symbols in Figure 63). The majority of these were located off southwestern Nova Scotia and in the eastern Gulf of Maine. No NOAA records occurred in the Laurentian Channel where the model predicted a highest presence probability of sea pens. Several records occurred in deeper waters off the shelf in an area considered as extrapolated by the model.

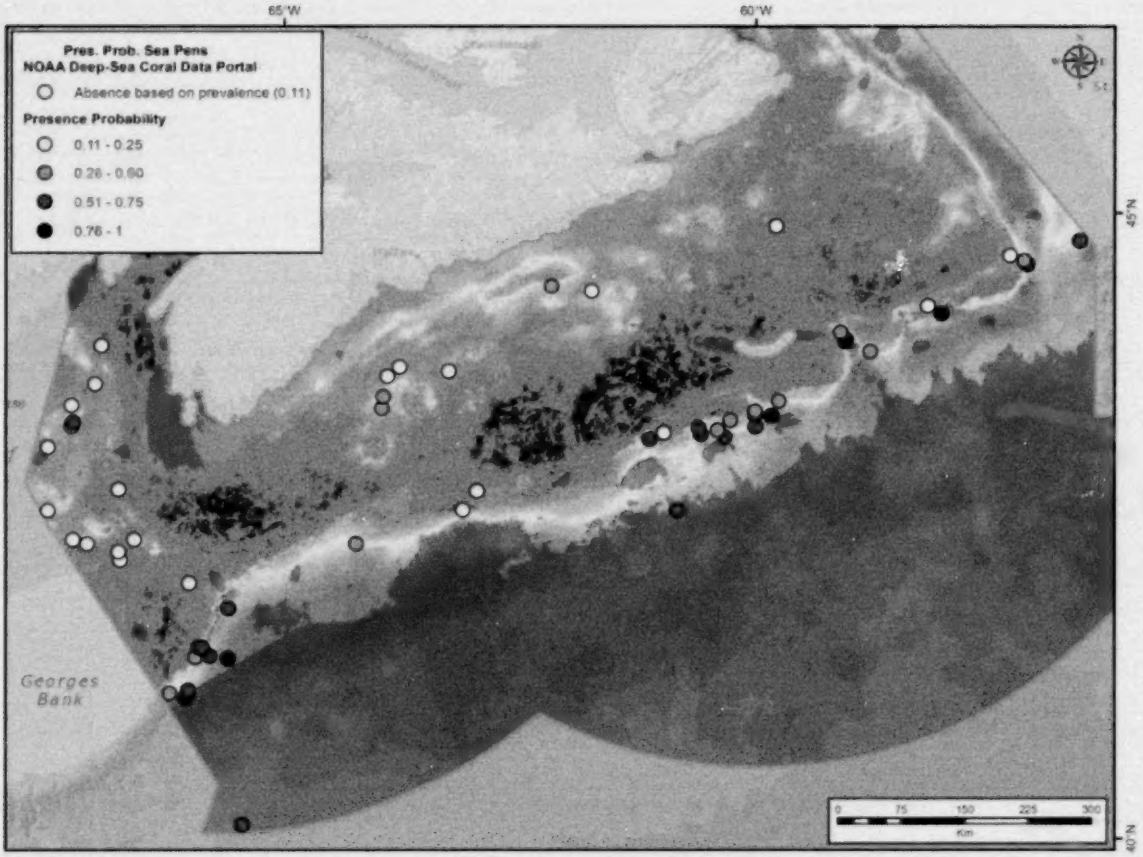

Figure 63. Validation of sea pen presence probability from Model 3 using independent data. Presence probability values were extracted to the location of sea pen records from the NOAA Deep-Sea Coral Data Portal.

Prediction of Sea Pen Biomass Using Random Forest

The accuracy measures of the regression random forest model on mean sea pen biomass per grid cell from DFO multispecies trawl surveys are presented in Table 18. The highest R^2 value was 0.814, while the average was 0.518 ± 0.301 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.018 ± 0.018 SD. This model explained a relatively high percentage of variance in the biomass data (average = 18.41% ± 2.48 SD).

Figures 64 and 65 show the predicted biomass surface of sea pens. The majority of the spatial extent was predicted to have low (0 – 1.21 kg) sea pen biomass. The Laurentian Channel has the highest predicted biomass of sea pens (up to 18.14 kg), coinciding with the location of the highest mean biomass values from the multispecies trawl surveys (Figure 65). Similar to the presence-absence models, much of the Laurentian Channel is predicted to have a moderate biomass of sea pens, despite there being no data records from there. The northeast portion of the Laurentian Channel is considered an area of extrapolation.

Table 18. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of average sea pen biomass (kg) per grid cell recorded from DFO multispecies trawl surveys in the Maritimes Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.641	0.177	0.006	18.19
2	0.253	0.545	0.018	22.32
3	0.781	0.491	0.016	19.86
4	0.509	0.475	0.016	20.19
5	0.814	0.142	0.005	18.26
6	0.694	2.014	0.066	19.00
7	0.673	0.290	0.009	15.62
8	0.745	0.180	0.006	13.42
9	8.478×10^{-5}	0.392	0.013	19.54
10	0.071	0.738	0.024	17.72
Mean	**0.518**	**0.544**	**0.018**	**18.41**
SD	**0.301**	**0.550**	**0.018**	**2.48**

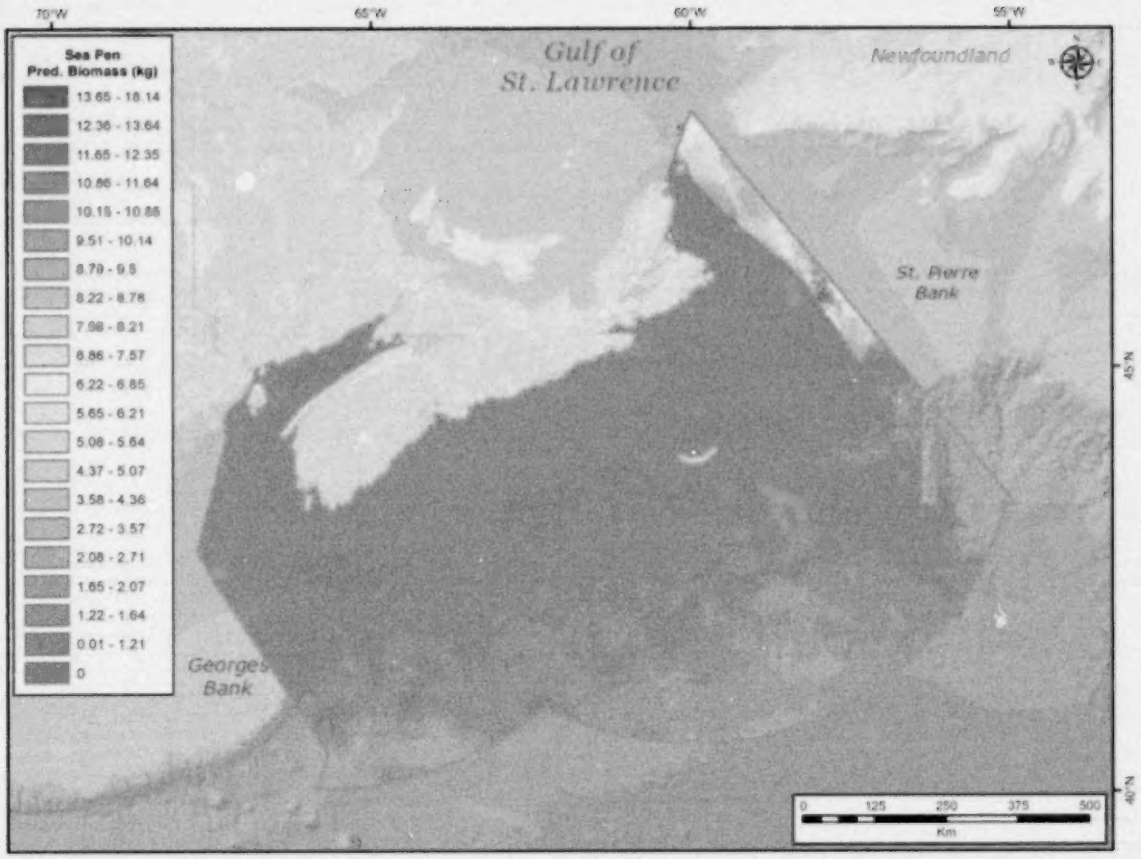

Figure 64. Predictions of biomass (kg) per grid cell of sea pens from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2002 and 2014.

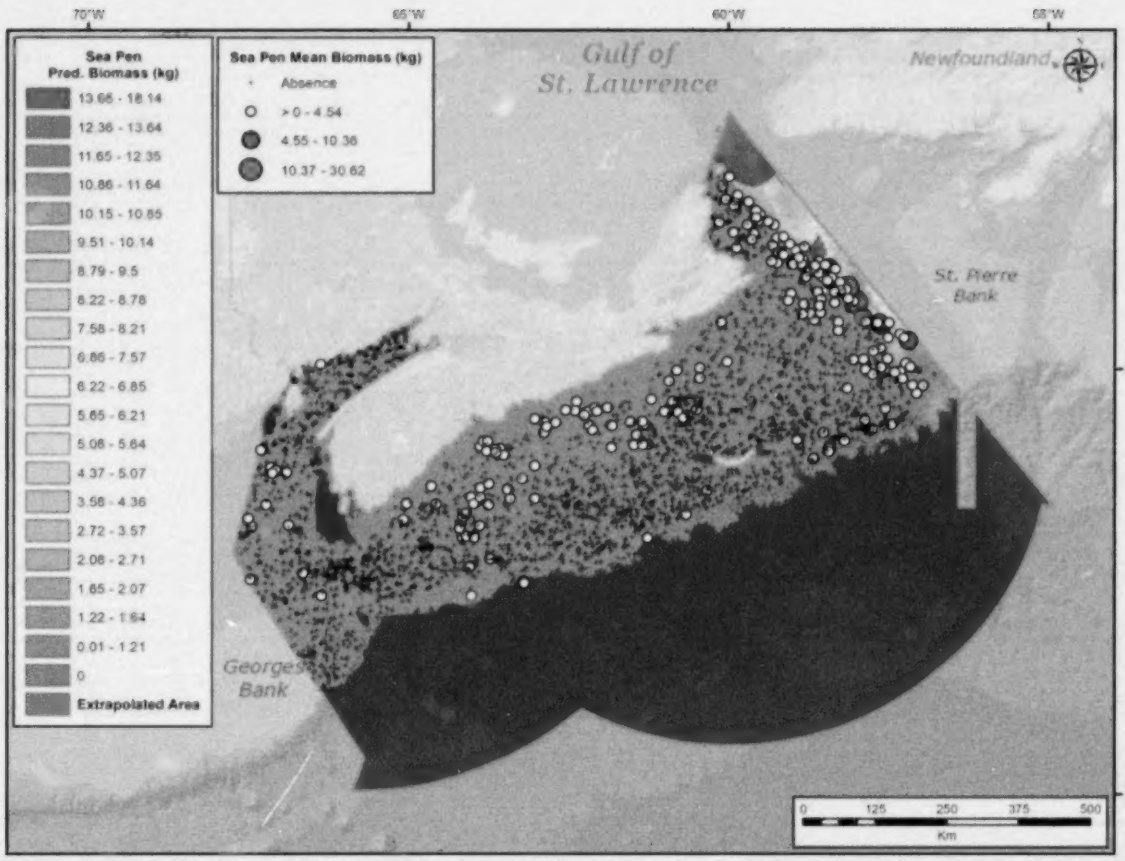

Figure 65. Predictions of biomass (kg) per grid cell of sea pens from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2002 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting sea pen biomass are shown in Figure 66. Bottom Salinity Average Range was the most important variable in the model. Prior to spatial interpolation, this variable displayed a right-skewed distribution with outlying data in the upper range (Beazley et al., in prep). Examination of the Q-Q plot revealed a strong spatial pattern to data points over- and under-predicted by a normal distribution, with over-predicted points located mainly off the central and western coast of Nova Scotia and in the deepest regions of the study extent, and under-predicted points located in the centre of the study extent just beyond the shelf break and in the Laurentian Channel. Bottom Salinity Average Range was followed closely by Bottom Shear Average Range and Surface Temperature Average Maximum. The partial dependence of sea pen biomass on the top 6 most important variables is shown in Figure 67. Predicted biomass was highest at Bottom Salinity Average Range values less than 0.5. These values coincided with both over- and under-predicted data points in the Laurentian Channel and in deep waters beyond the shelf. The fit between predicted and observed

values for this variable was fair, with slight over-prediction of Bottom Salinity Average Range values < 0.5. Some points could therefore be predicted higher than their true values and slightly outside the range of highest predicted biomass identified in the partial plot.

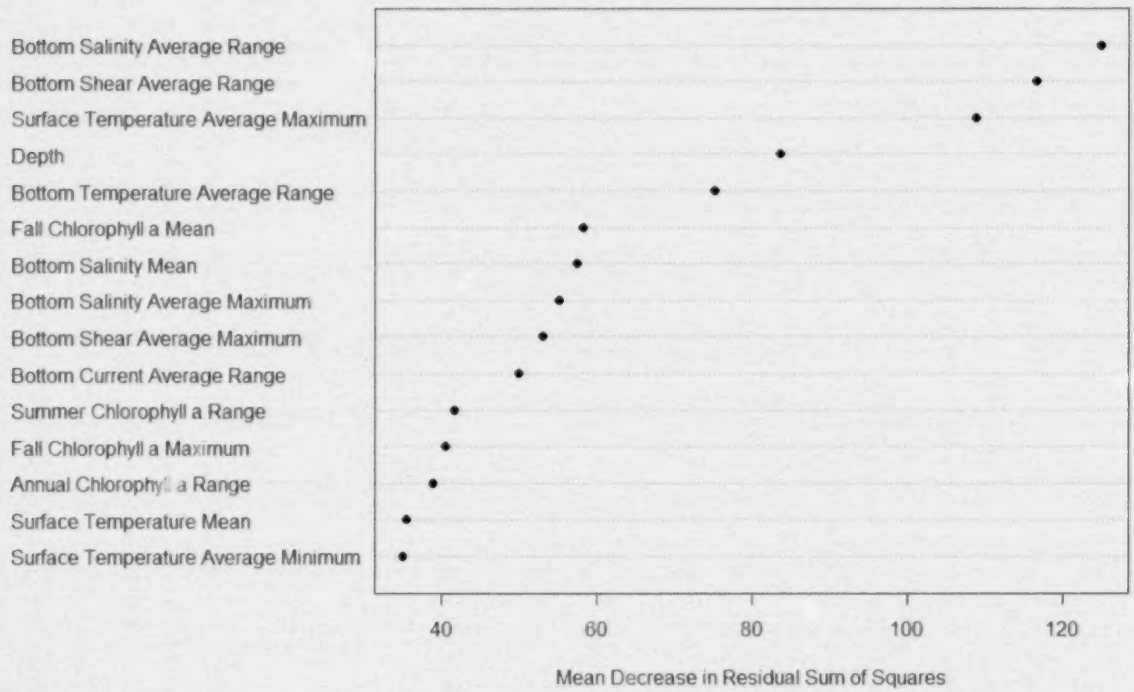

Figure 66. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on sea pen mean biomass data averaged per grid cell. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

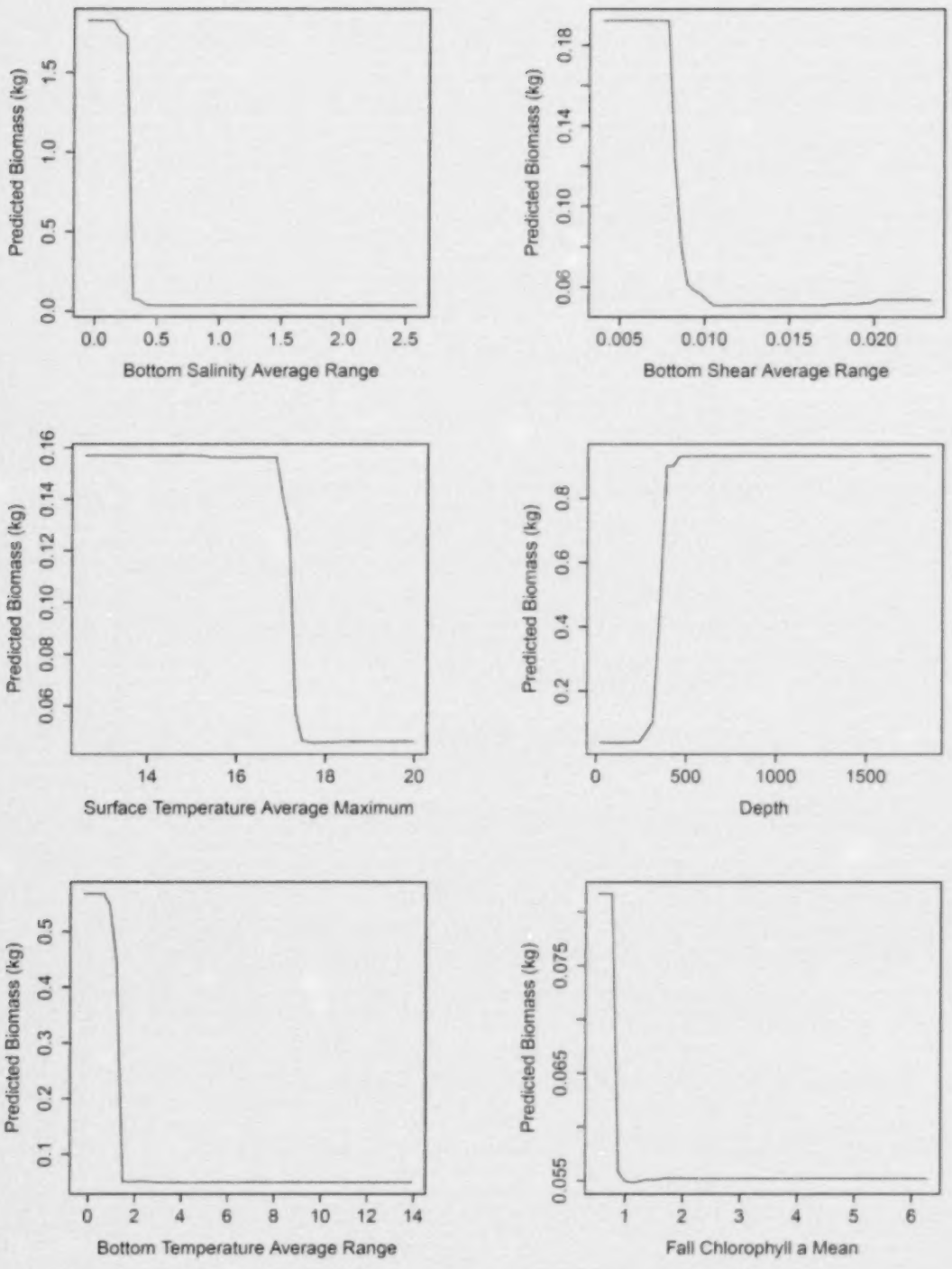

Figure 67. Partial dependence plots of the top six predictors from the random forest model of sea pen biomass data collected within the Maritimes Region, ordered left to right from the top. Predicted biomass (kg) is shown on the *y*-axis.

Large Gorgonian Corals

Data Sources and Distribution

Figure 68 shows the distribution of available large gorgonian records in the Maritimes Region. There was relatively good congruence in the spatial distribution of records originating from the different data sources. DFO multispecies trawl survey records were concentrated on the banks of the eastern Scotian Shelf and along the slope. Notably, several records occurred on the shallow banks around Cape Breton where large gorgonian occurrences from the Gass (2002) and Breeze (1997) were reported. The scientific survey and NOAA records were concentrated along the slopes and in the Northeast Channel and eastern Gulf of Maine.

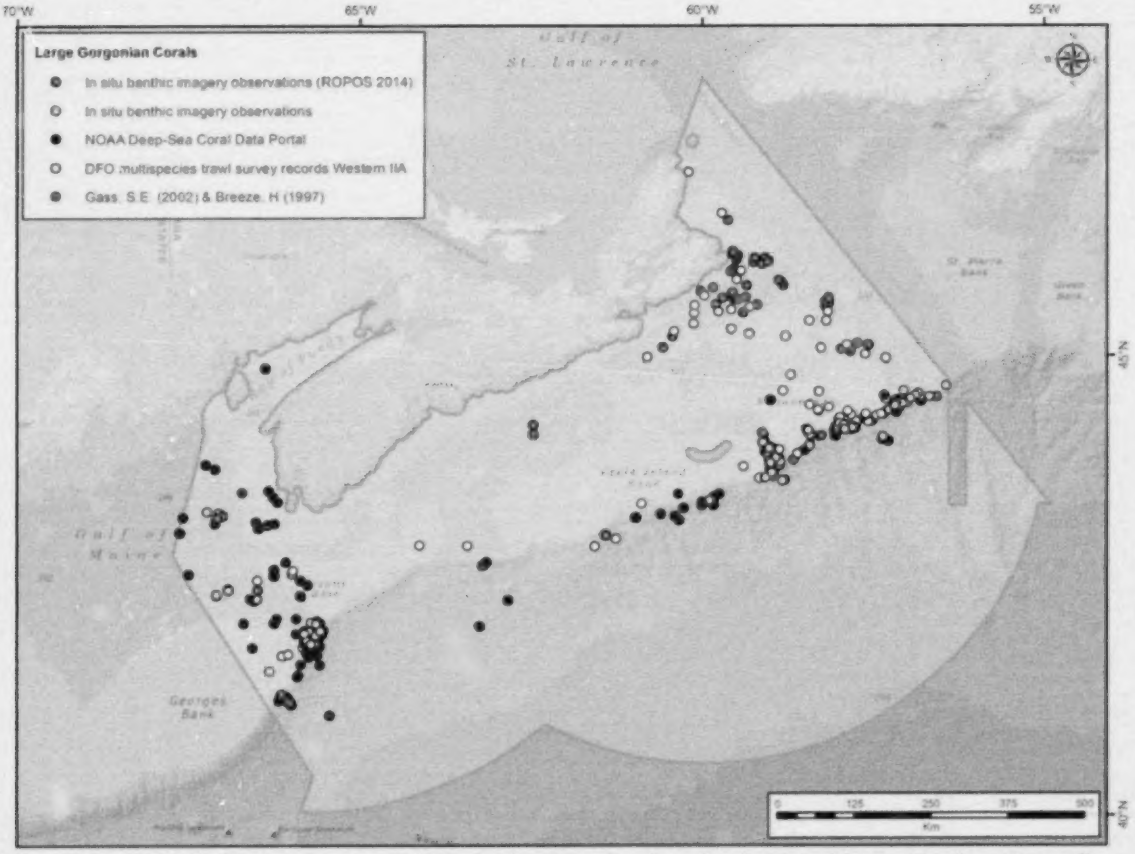

Figure 68. Available large gorgonian coral presence data in the Maritimes Region from Gass (2002) and Breeze et al. (1997), scientific missions, the NOAA Deep-Sea Coral Data Portal, and DFO multispecies research vessel surveys.

Initial random forest models of large gorgonians were run using only catch data originating from DFO multispecies trawl surveys (Western IIA gear). This data was collected over a period of 13 years from 2002 to 2015 (Table 19). This dataset consisted of 72 presence and 2313 absence records (Figure 69). Absence records were distributed relatively evenly across the Scotian Shelf and Bay of Fundy. The highest mean biomass record (up to 54.20 kg) occurred on the slopes between Haldimand Canyon and the *Lophelia* Coral Conservation Area in the Stone Fence. The second highest mean biomass record (27.11 kg) occurred in the Northeast Channel.

Table 19. Number of presence and absence records of large gorgonian coral catch recorded from DFO multispecies trawl surveys conducted between 2002 and 2015 in the Maritimes Region.

Year	Total number of presences	Total number of absences
2002	1	96
2003	7	177
2005	8	260
2006	3	175
2007	2	173
2008	1	62
2009	8	191
2010	5	297
2011	8	250
2012	10	205
2013	11	205
2014	3	194
2015	5	28

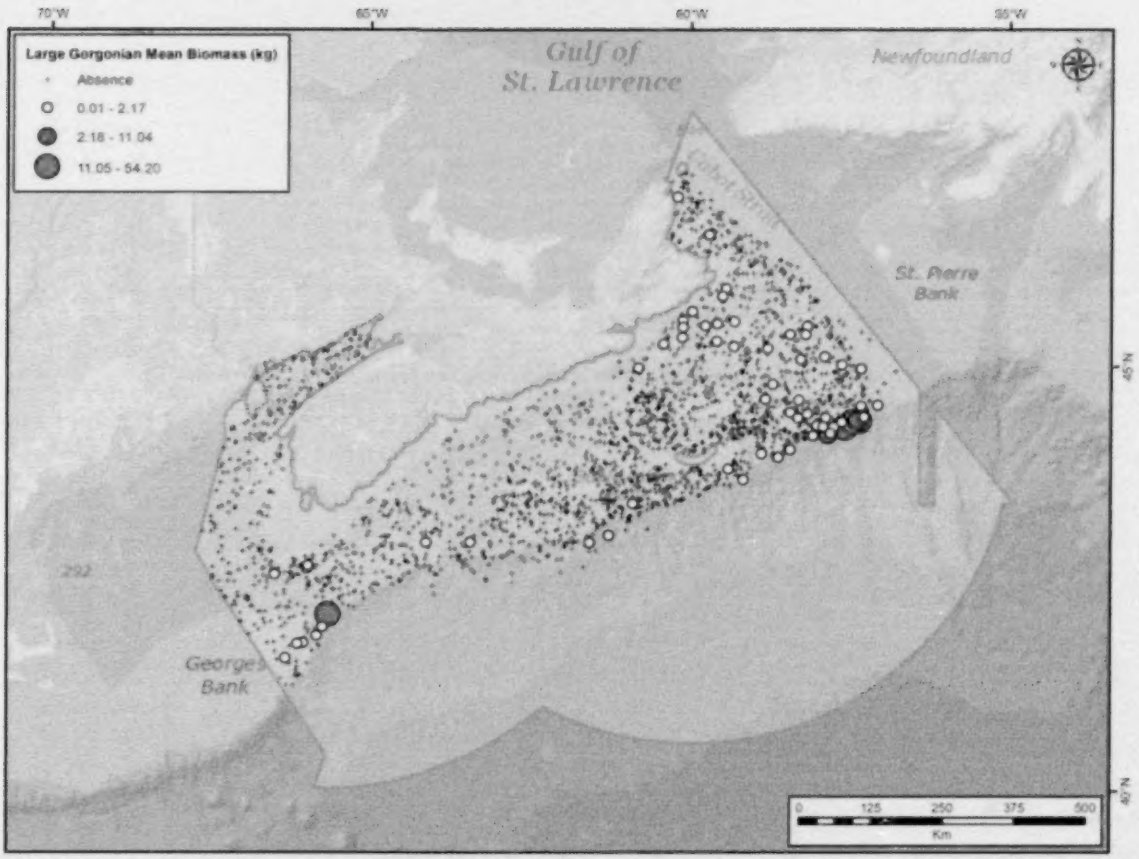

Figure 69. Mean biomass (kg) per grid cell of large gorgonian coral catch recorded from DFO multispecies trawl surveys from 2002 to 2015 within the Maritimes Region. Also shown are absence records from the same surveys.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity, and specificity) for the random forest model on balanced species prevalence (199 presences and 199 absences; Model 1) are presented in Table 20. The highest mean AUC of 0.861 was associated with Model Run 1 and is therefore considered the optimal model for the prediction of the large gorgonian coral response data. The sensitivity and specificity measures of this model run were 0.750 and 0.806, respectively. The confusion matrix of the optimal model is also presented in Table 2. Class error for both the presence and absence classes was somewhat moderate (0.194 and 0.250, respectively), and was slightly higher for the presence class.

The presence probability prediction surface of large gorgonian corals is presented in Figure 70. The highest predictions of presence probability occurred on the eastern Scotian Slope and Scotian Shelf on Banquereau and Misaine Banks. High presence probability of large gorgonian corals was also predicted to occur off the coast of Cape Breton. The Northeast Channel area had a moderate to high presence probability of large gorgonians. These areas corresponded well with the spatial distribution of presence records (see Figure 71) and areas of high presence probability do not appear to be grossly extrapolated beyond presence locations.

Table 20. Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of large gorgonian corals within the Maritimes Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 1) which is considered the optimal model for predicting the presence probability of large gorgonians in the region.

Model Run	AUC	Sensitivity	Specificity
1	**0.861**	**0.750**	**0.806**
2	0.730	0.653	0.694
3	0.787	0.722	0.653
4	0.849	0.750	0.750
5	0.859	0.833	0.778
6	0.814	0.694	0.708
7	0.783	0.708	0.708
8	0.821	0.750	0.764
9	0.814	0.722	0.736
10	0.834	0.750	0.736
Mean	**0.815**	**0.733**	**0.733**
SD	**0.040**	**0.047**	**0.044**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	58	14	72	0.194
Presence	18	54	72	0.250

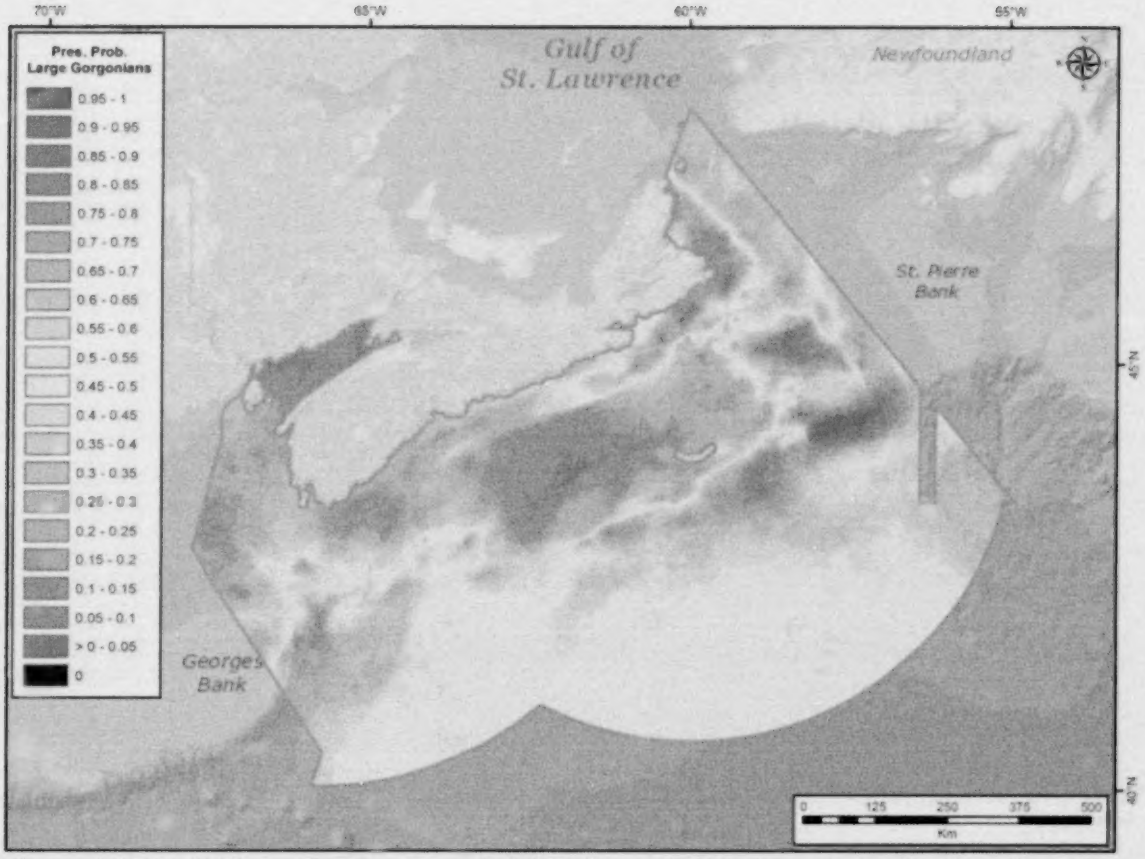

Figure 70. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of large gorgonian coral presence and absence data collected from DFO multispecies trawl surveys in the Maritimes Region between 2002 and 2015.

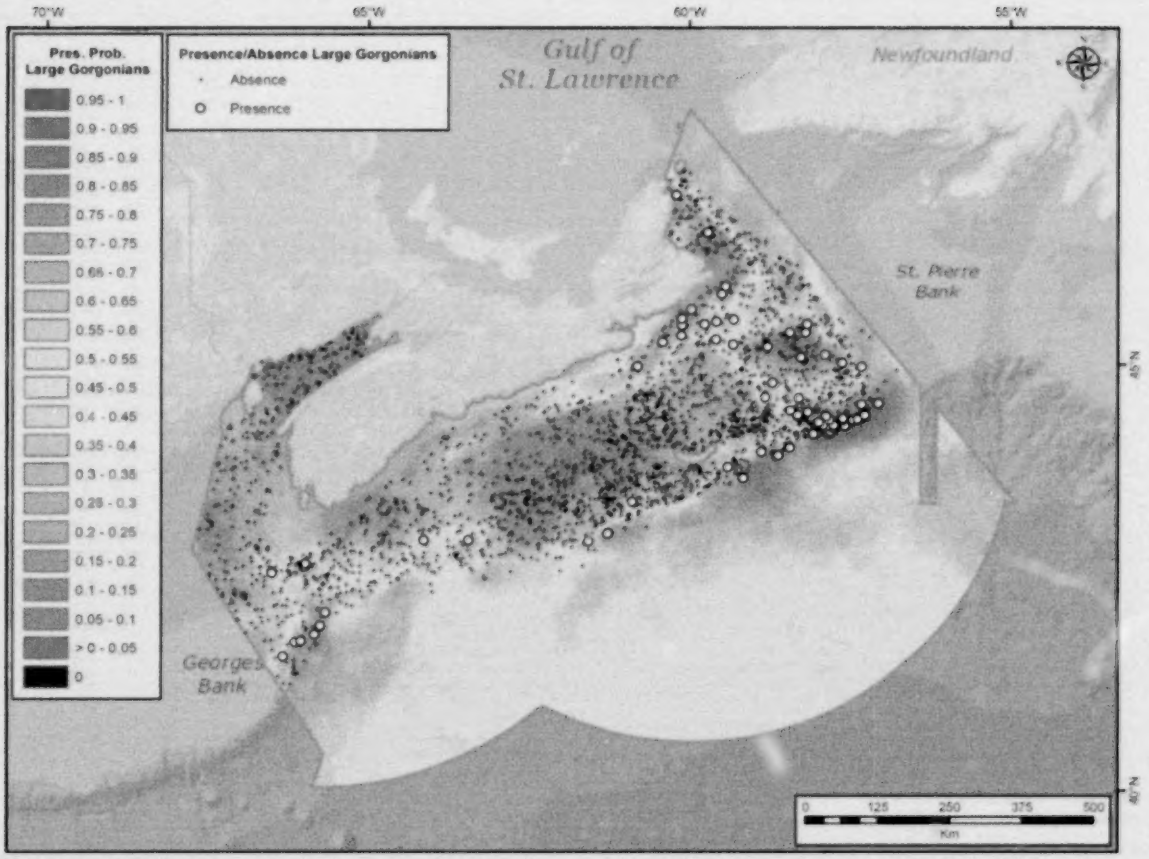

Figure 71. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of large gorgonian coral presence and absence data recorded from DFO multispecies trawl surveys in the Maritimes Region between 2002 and 2015.

The actual presence and absence data observations used in the optimal model run of Model 1 (72 presences and 72 absences; Figure 72) showed some slight spatial bias across the study area. Despite there being absence records on the eastern Scotian Shelf on Misaine and Banquereau Banks, very few were selected from these areas during the random down-sampling of the data prior to modelling. This likely caused a slight over-extension of high predicted probabilities in these areas. Areas of extrapolation of the balanced prevalence random forest model on large gorgonian corals are also shown in Figure 72. All deep water beyond the Scotian Shelf is considered extrapolated area. Large pockets of extrapolated area occurred off southwestern Nova Scotia and northeast tip of Cape Breton. Smaller pockets of extrapolated area occurred across the Scotian Shelf. Areas of extrapolation do not appear to greatly overlap with areas of high predicted presence probability in the shallow portion of the study area except for along the east coast of Cape Breton.

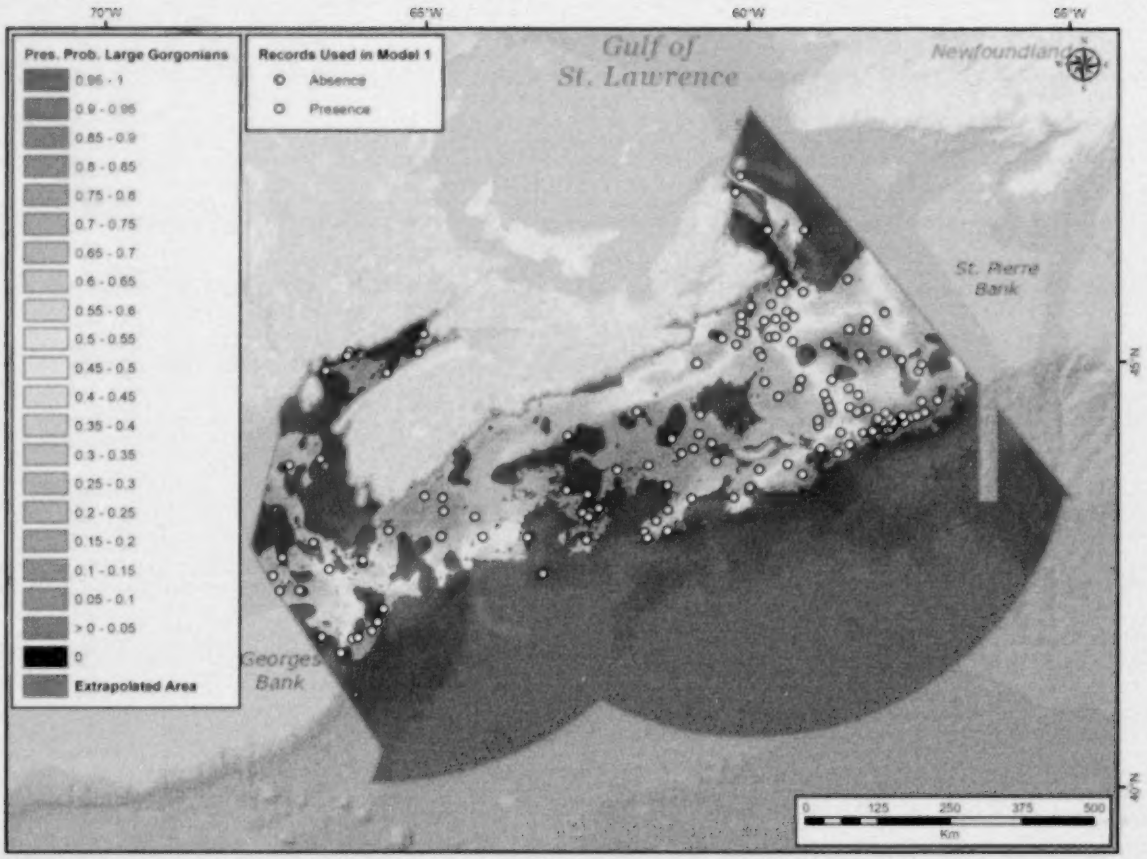

Figure 72. Map of the 144 data observations (72 presences and 72 absences) of large gorgonian corals used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of large gorgonians and the areas of model extrapolation (i.e. areas where at least one environmental predictor variable is outside of the sampled range.

Of all 66 environmental predictor variables used in the model, Surface Current Mean was the most important for the classification of the large gorgonian presence and absence data (Figure 73). Prior to spatial interpolation, this variable displayed a slightly right-skewed distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed a spatial pattern to data points over- and under-predicted by a normal distribution, with over-predicted points located in the centre of the study extent and off the central and western coast of Nova Scotia, and under-predicted points located mainly in Bay of Fundy and in the deepest regions of the study extent. This variable was followed more distantly in terms of its Mean Decrease in Gini Value by Maximum Average Spring Mixed Layer Depth (m). Surface current and mixed layer depth variables ranked high in the model. Partial dependence plots for the top 6 predictor variables are shown in Figure 74. Presence probability of large gorgonians was highest at Surface Current

Mean values greater than 0.10 m s^{-1}. Values in this range coincided with both over- and under-predicted data points located off Cape Breton and in the deepest portion of the study extent. These values are not of concern however, as the fit between predicted and observed values was excellent with only slight over-prediction of Surface Current Mean values of 0.10 m s^{-1} and greater. Presence probability was greatest at the highest values along the other surface current, temperature, and mixed layer depth variables.

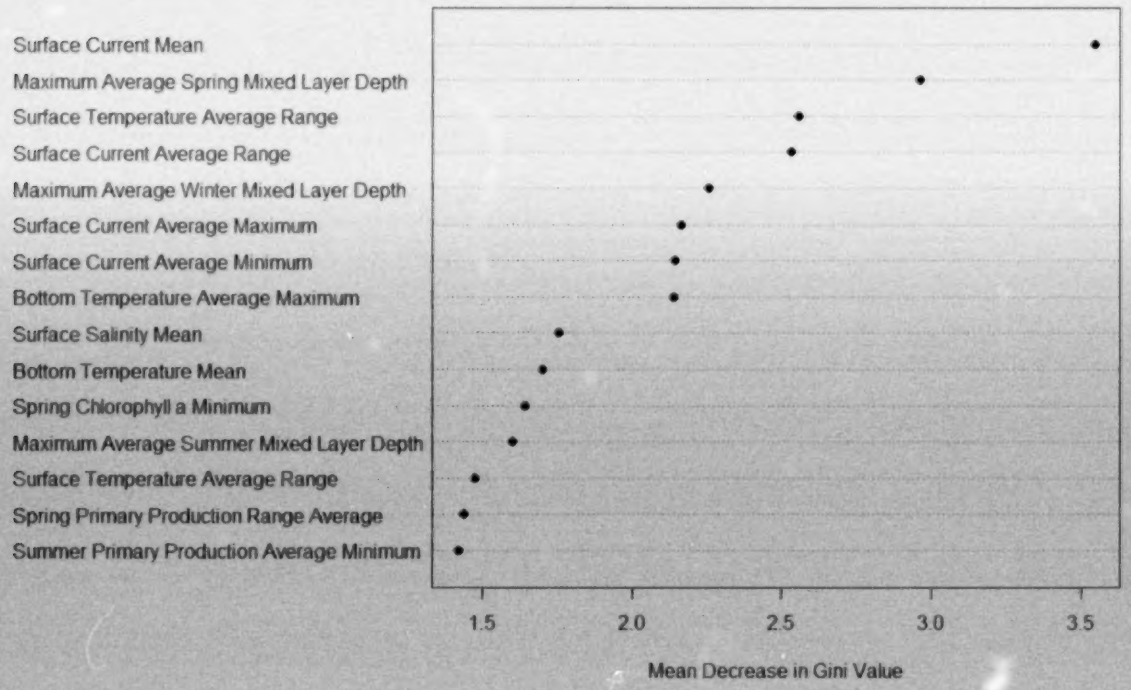

Figure 73. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting large gorgonian coral presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

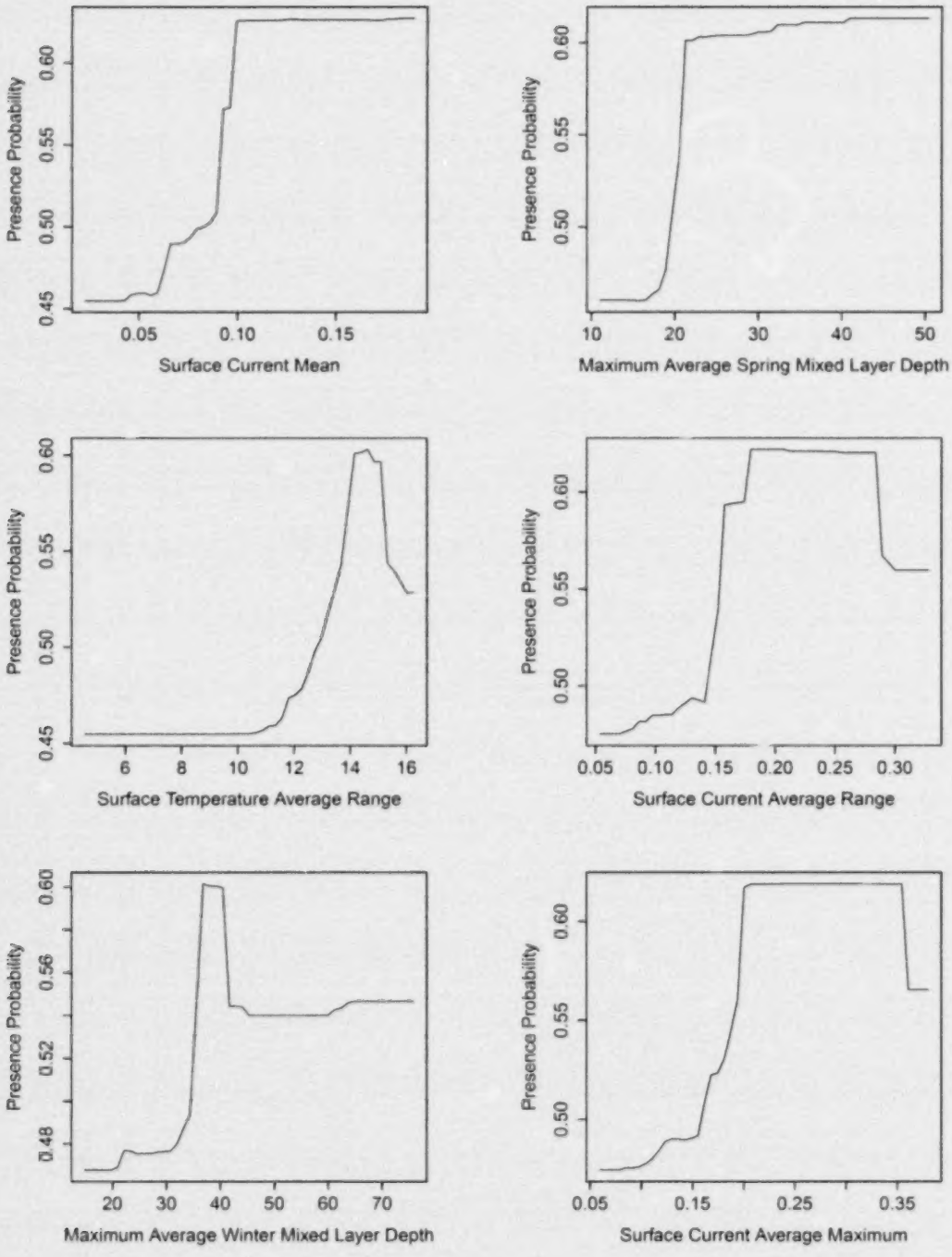

Figure 74. Partial dependence plots of the top six predictors from the optimal random forest model of large gorgonian coral presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 2 - Unbalanced Data and Threshold Equal to Species Prevalence

Table 21 shows the accuracy measures for the random forest model on all large gorgonian coral presence and absence data (72 presences and 2313 absences; Model 2) and a threshold equal to species prevalence (0.03). The average AUC calculated from Model 2 was 0.797, lower than that of Model 1. Class error of the absence class was similar to that of Model 1 (0.215 compared to 0.194 from Model 1), however class error for the presence class was slightly higher (0.292 compared to 0.250 from Model 1). Sensitivity and specificity measures of Model 2 were lower than that of Model 1.

The surface of predicted presence probability of large gorgonian corals generated from Model 2 is shown in Figure 75. Only the slope area between Haldimand Canyon and the Stone Fence was predicted to have moderate to high presence probability of large gorgonians. Much of central Scotian Shelf and Bay of Fundy was predicted to have zero or low presence probability of large gorgonians. The model does not appear to predict areas of presence far beyond the location of presence records (Figure 76), likely due to the inclusion of absences records in the model where presence records also occurred. Areas of extrapolation of Model 2 are presented in Figure 77. Areas of extrapolation do not overlap with areas of high presence probability in the shallow portion of the study area.

Table 21. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of large gorgonian corals within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.715		Absence	Presence				
2	0.924	Absence	1815	498	2313	0.215	0.708	0.785
3	0.891	Presence	21	51	72	0.292		
4	0.589							
5	0.763							
6	0.851							
7	0.795							
8	0.869							
9	0.714							
10	0.854							
Mean	**0.797**							
SD	**0.102**							

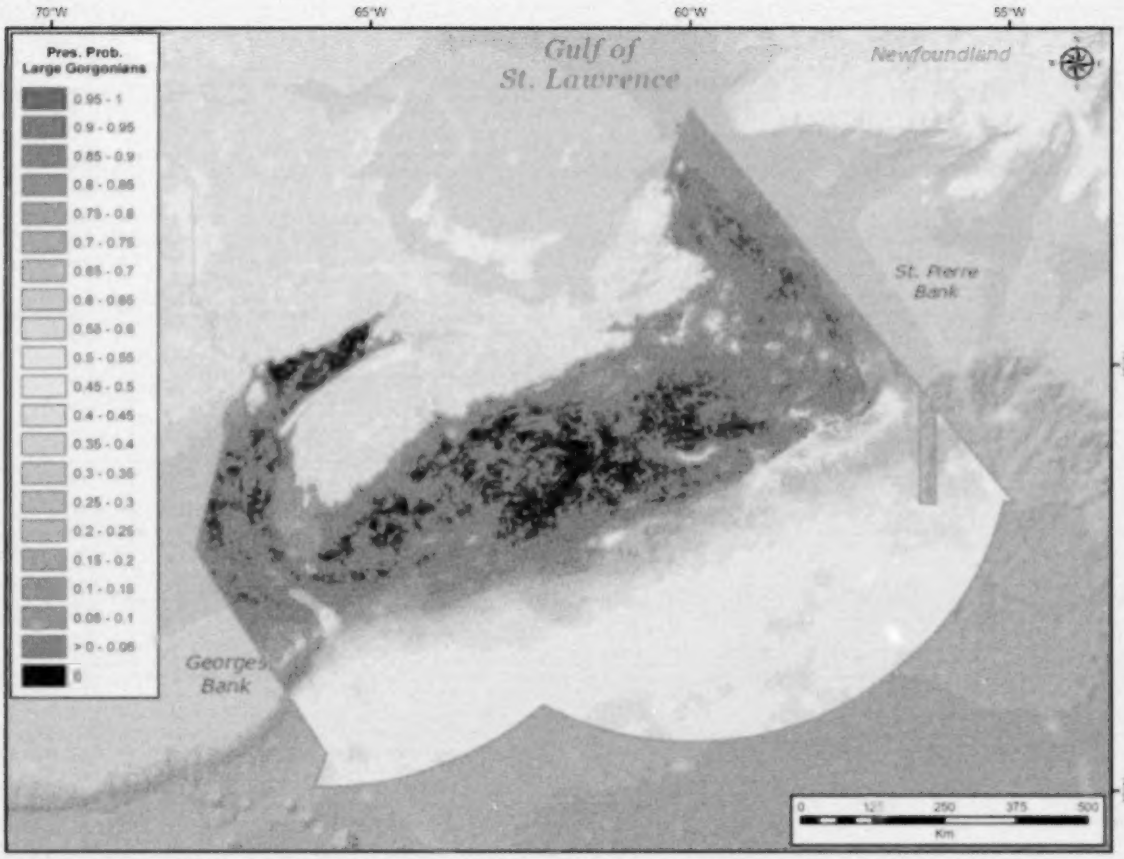

Figure 75. Predictions of presence probability (Pres. Prob.) of large gorgonian corals based on a random forest model on unbalanced presence and absence large gorgonian catch data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2002 and 2015.

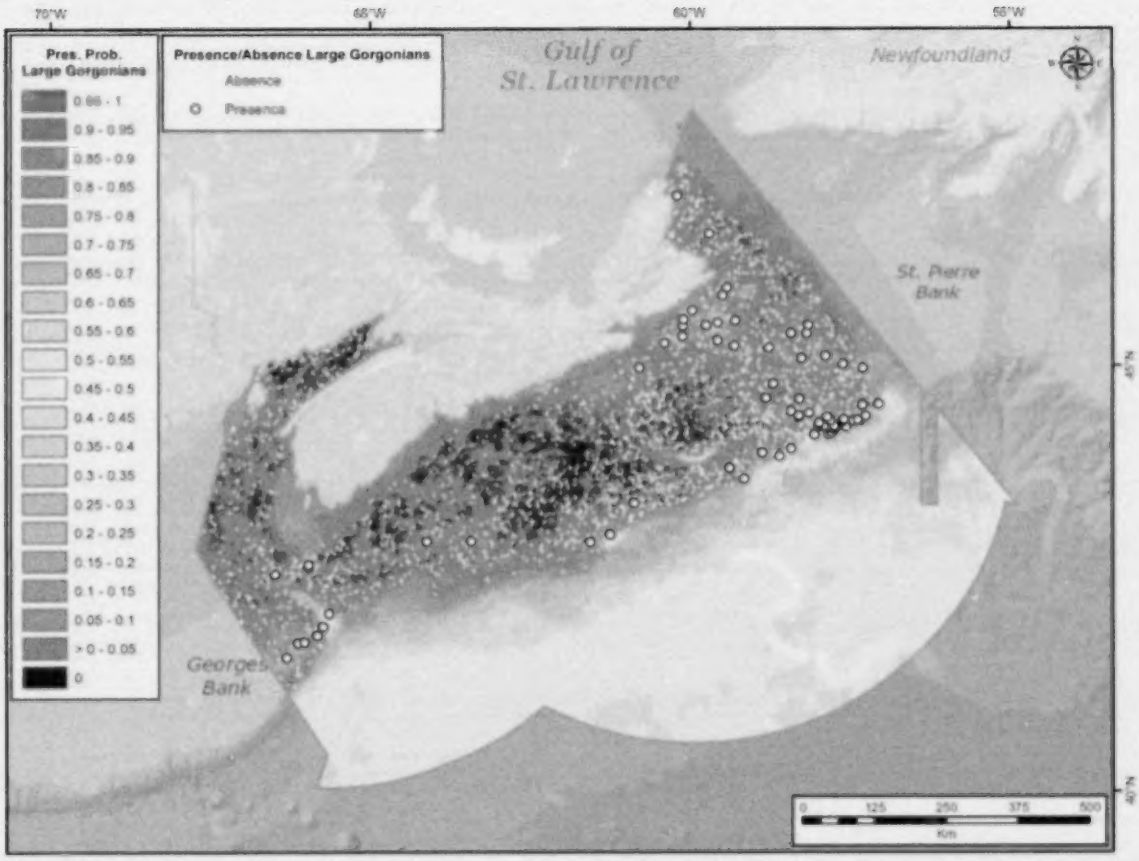

Figure 76. Presence and absence observations and predictions of presence probability (Pres. Prob.) of large gorgonian corals based on a random forest model on unbalanced presence and absence large gorgonian catch data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2002 and 2015.

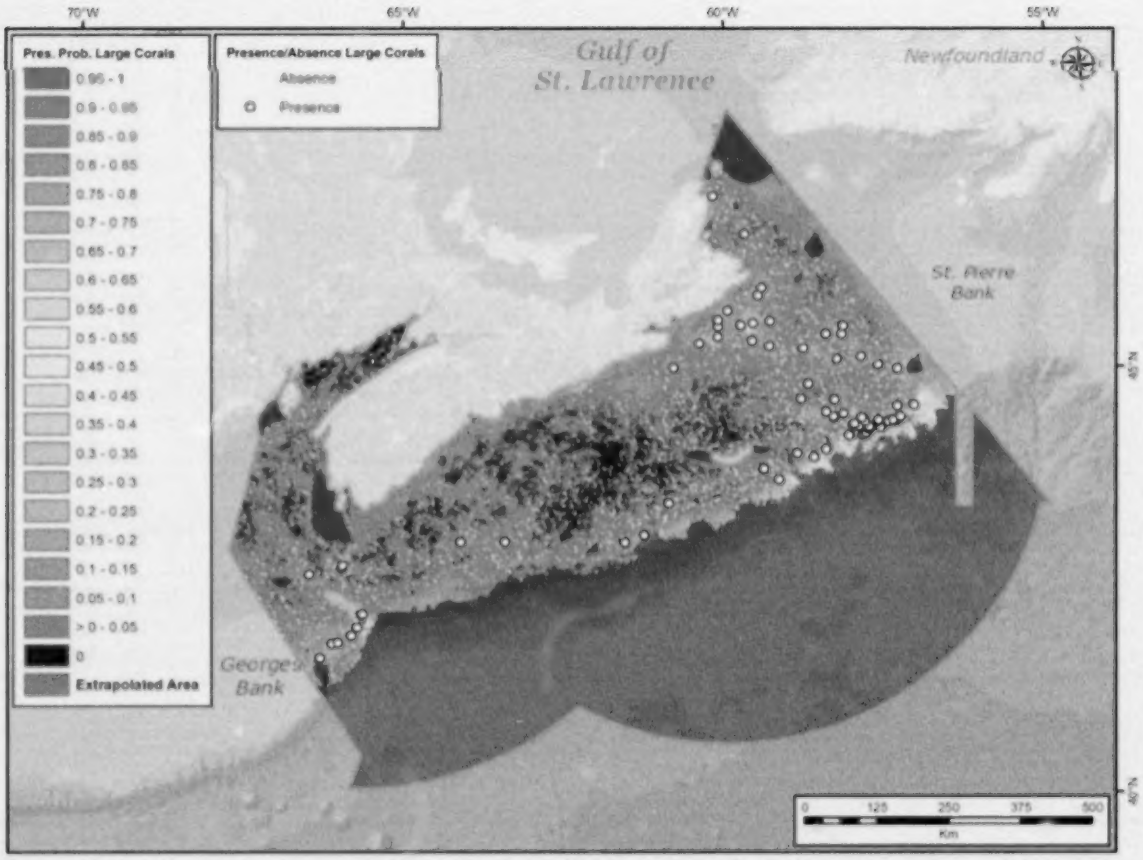

Figure 77. Areas of extrapolation of the random forest model on unbalanced presence and absence large gorgonian catch data collected from the Maritimes Region between 2002 and 2015. Also shown are the large gorgonian presence and absence observations and predictions of presence probability (Pres. Prob.).

The order of importance of the environmental predictor variables in Model 2 was slightly different from that of Model 1 (Figure 78). Slope was the most important variable in Model 2 compared to Surface Current Mean in Model 1. Slope was not among the top 15 most important variables in Model 1. Partial dependence of large gorgonian presence and absence data on the top 6 predictor variables is shown in Figure 79. Slope was followed very distantly by Surface Current Mean and the remaining variables. A small peak in presence probability at 3° occurred along the Slope gradient. Presence probability then rapidly increased at 5° and plateaued. Similar to Model 1, large gorgonian presence probability was highest at higher surface current values.

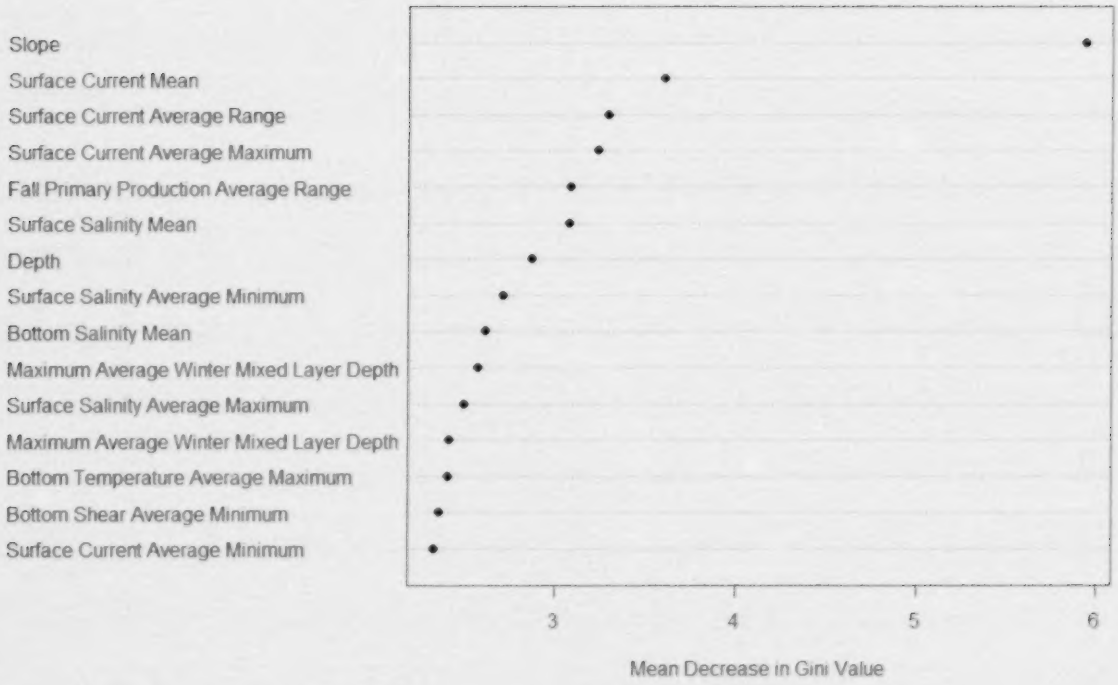

Figure 78. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced large gorgonian coral presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

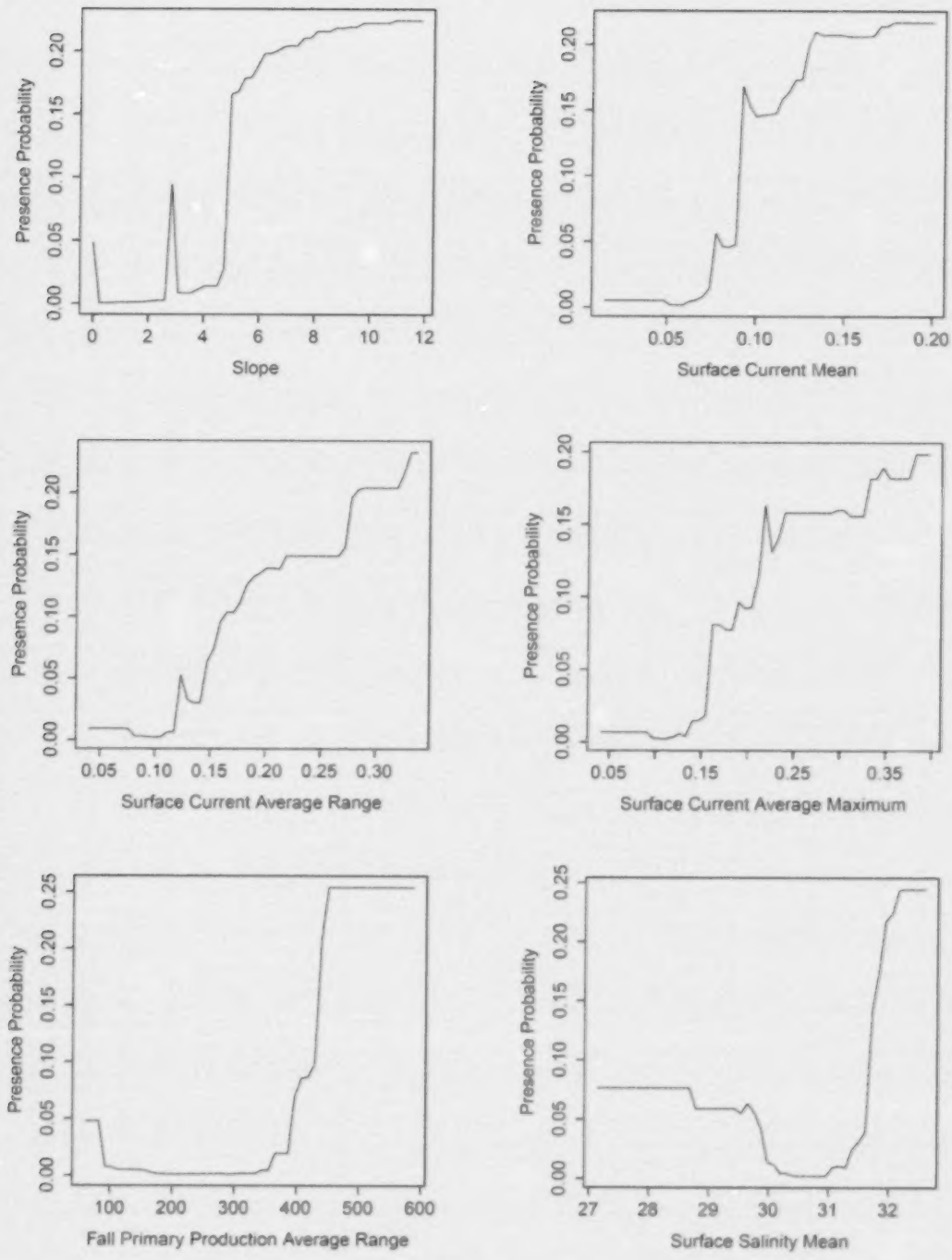
Presence Probability
0.00 0.05 0.10 0.15 0.20
0 2 4 6 8 10 12
Slope
Presence Probability
0.00 0.05 0.10 0.15 0.20
0.05 0.10 0.15 0.20
Surface Current Mean
Presence Probability
0.00 0.05 0.10 0.15 0.20
0.05 0.10 0.15 0.20 0.25 0.30
Surface Current Average Range
Presence Probability
0.00 0.05 0.10 0.15 0.20
0.05 0.15 0.25 0.35
Surface Current Average Maximum
Presence Probability
0.00 0.05 0.10 0.15 0.20 0.25
100 200 300 400 500 600
Fall Primary Production Average Range
Presence Probability
0.00 0.05 0.10 0.15 0.20 0.25
27 28 29 30 31 32
Surface Salinity Mean

Figure 79. Partial dependence plots of the top six predictors from the random forest model of large gorgonian coral unbalanced presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 3 – Addition of *In Situ* Benthic Imagery Observations

Given the low number of large gorgonian presence records in the Maritimes Region, the DFO multispecies trawl survey data were augmented with additional *in situ* benthic imagery records collected in the Maritimes Region between 1967 and 2011. A total of 155 additional presence records (Table 22) were added to the dataset after filtering the data so that only one presence record occurred per environmental grid cell. The combined dataset consisting of 227 presences and 2313 absences was remodelled (termed Model 3) using an unbalanced design and a threshold equal to species prevalence (0.09). The accuracy measures for random forest Model 3 are shown in Table 23. The average AUC computed from 10-fold cross validation was 0.928 ± 0.033 SD, the highest of all three models. Class error for the presence and absence classes was the lowest of all three models, while sensitivity and specificity were the highest.

Table 22. Number of *in situ* benthic imagery observations of large gorgonians collected from scientific surveys conducted between 1967 and 2011 in the Maritimes Region.

Year	Gear	Total number of presences
1967	NRCan Drop Camera	1
1997	Campod	1
1999	Campod	1
2000	NRCan Drop Camera	1
2000	Campod	11
2001	Campod	24
2001	ROPOS	3
2001	ROPOS (Martha Black)	1
2002	Campod	8
2003	Campod	10
2005	Campod	15
2006	ROPOS	21
2006	DSIS ROV	2
2007	ROPOS	28
2008	NRCan Drop Camera	1
2008	Campod	23
2011	Campod	4

The additional presence records expanded the area of high presence probability along the eastern Scotian Slope (Figure 80). The Gully MPA and area directly south of it showed a much higher presence probability compared to Models 1 and 2. Shortland and Haldimand canyons off Banquereau Bank and the slope in between also had an increased probability of occurrence of large gorgonian corals, as well as the Northeast Channel on the western Scotian Shelf. These areas of higher presence probability along the slope corresponded well with the location of presence records from the DFO and NRCan scientific surveys (Figure 81). The area of extrapolation along the slope off eastern Scotian Shelf is reduced with the addition of science survey records there (see Figure 82). Figure 83 depicts the classification of large gorgonian presence probability into presence and absence categories based on the prevalence threshold of 0.09. In this map, all presence probability values generated from Model 3 that were greater than 0.09 were classified as presence, while values less than 0.09 were classed as absence. The majority of the slope and deep-water channels were classified as presence of large gorgonians. The large area southwest of Nova Scotia avoided by trawl surveys due to hard bottom is also classified as presence for large gorgonians.

Table 23. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of large gorgonian corals from DFO multispecies trawl survey records and scientific surveys conducted within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.934		Absence	Presence				
2	0.882	Absence	2063	250	2313	0.108	0.833	0.892
3	0.910	Presence	38	189	227	0.167		
4	0.905							
5	0.935							
6	0.911							
7	0.956							
8	0.971							
9	0.981							
10	0.895							
Mean	**0.928**							
SD	**0.033**							

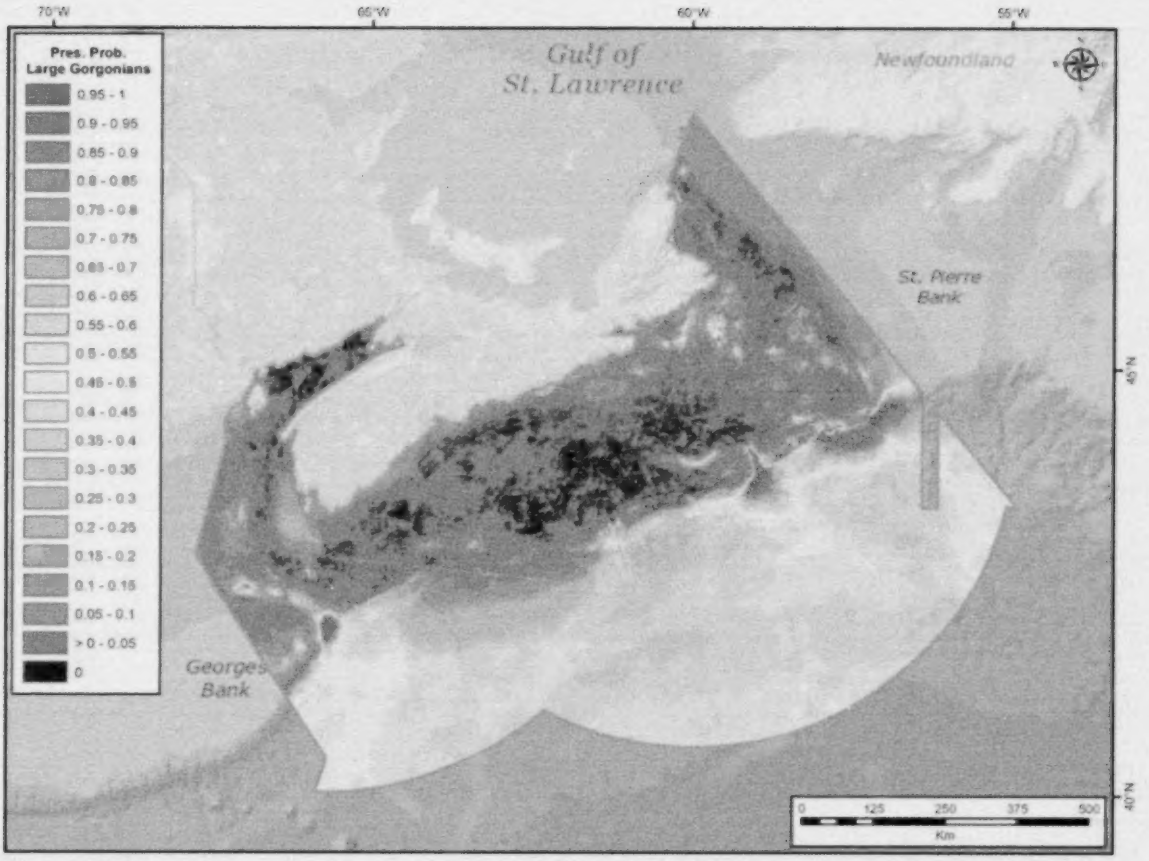

Figure 80. Predictions of presence probability (Pres. Prob.) of large gorgonian corals based on a random forest model on unbalanced presence and absence gorgonian catch data collected from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1967 and 2015.

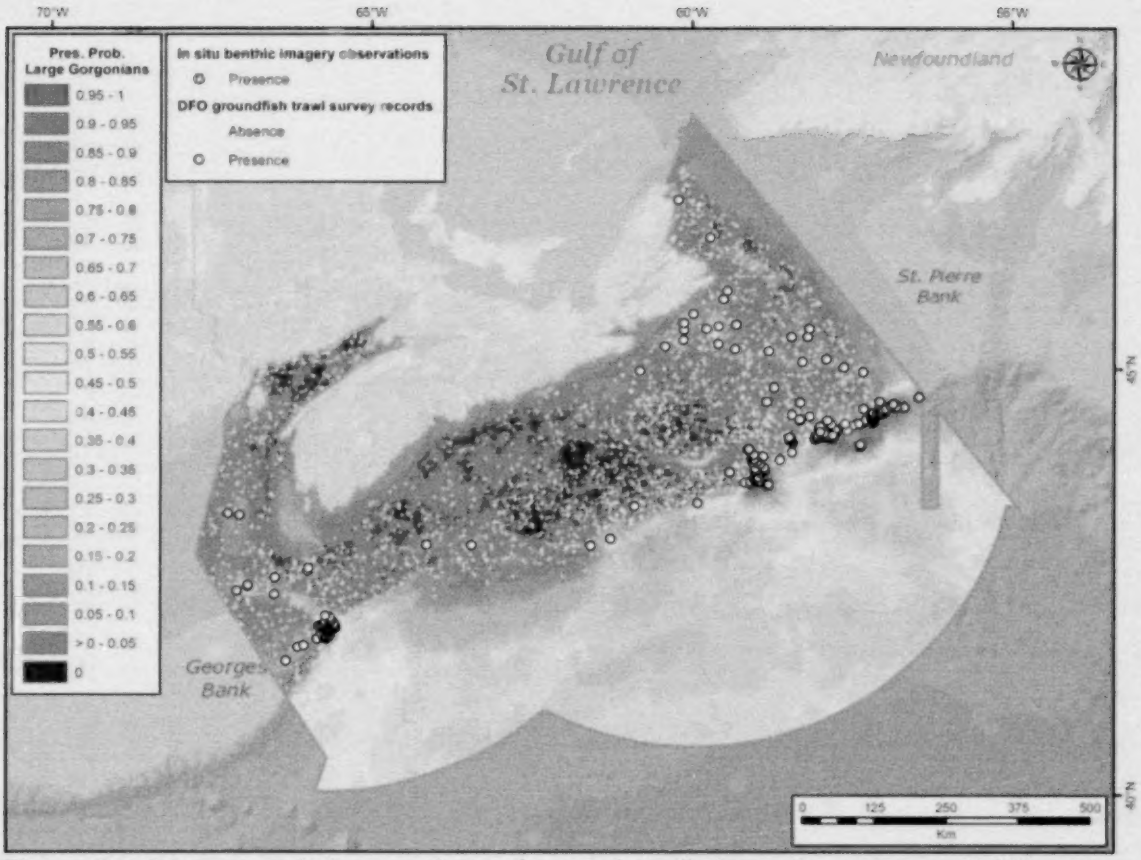

Figure 81. Presence and absence observations and predictions of presence probability (Pres. Prob.) of large gorgonian corals based on a random forest model on unbalanced presence and absence gorgonian catch data collected from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1967 and 2015.

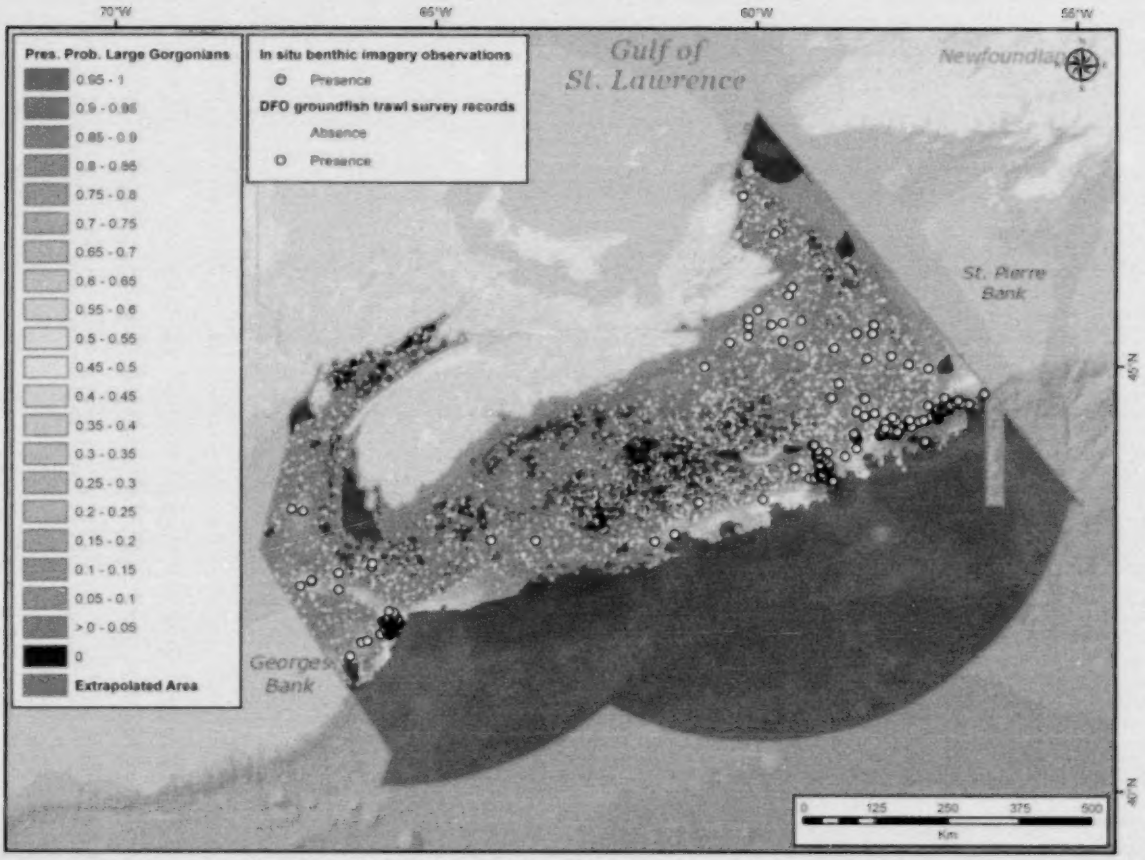

Figure 82. Areas of extrapolation of the random forest model on unbalanced presence and absence large gorgonian coral catch data from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1967 and 2015. Also shown are the presence and absence observations and predictions of presence probability (Pres. Prob.).

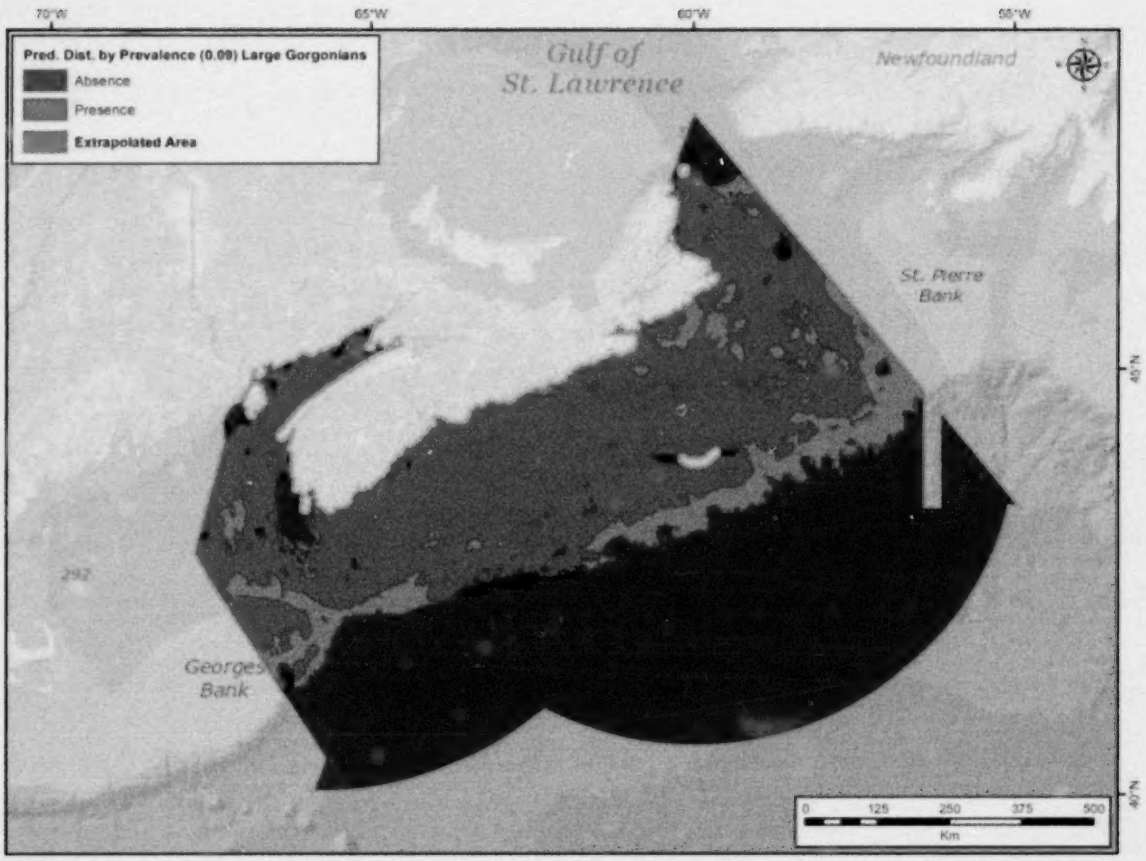

Figure 83. Predicted distribution (Pred. Dist.) of large gorgonian corals in the Maritimes Region based on the prevalence threshold of 0.09 of large gorgonian coral presence and absence data used in Model 3. Also shown are the areas of model extrapolation (grey polygon may appear red or blue).

Like in Model 2, the most important environmental predictor variable for the classification of the large gorgonian coral presence and absence data was Slope (Figure 84). This was followed closely by Depth, then more distantly by Surface Salinity Mean and the remaining variables in the model. Bottom and surface salinity, and surface current variables ranked high in this model. Partial dependence plots of the top 6 environmental variables are shown in Figure 85. Probability of presence of large gorgonians increased steadily along the Slope gradient and reached a plateau at ~7°. A similar pattern was shown along the Depth gradient, where probability of presence rapidly increased at ~250 m depth and plateaued at ~500 m. The bottom and surface salinity variables all showed the same pattern of high presence probability at the highest salinity levels.

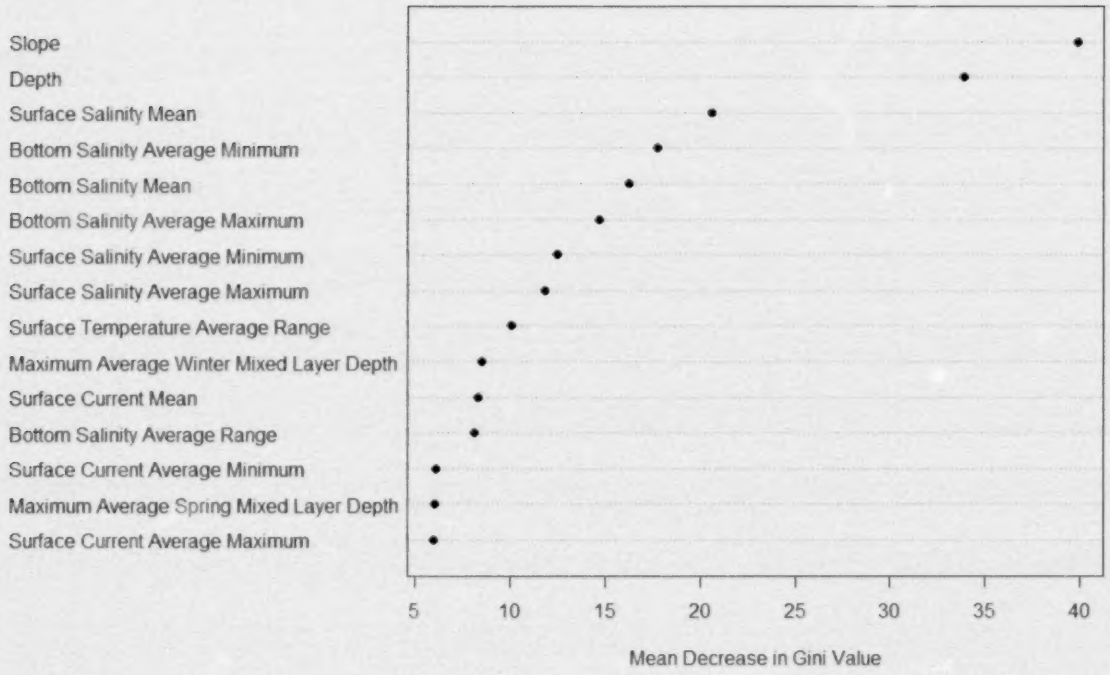

Figure 84. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced large gorgonian coral presence and absence data collected from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1967 and 2015. The higher the Mean Gini value the more important the variable is for predicting the response data.

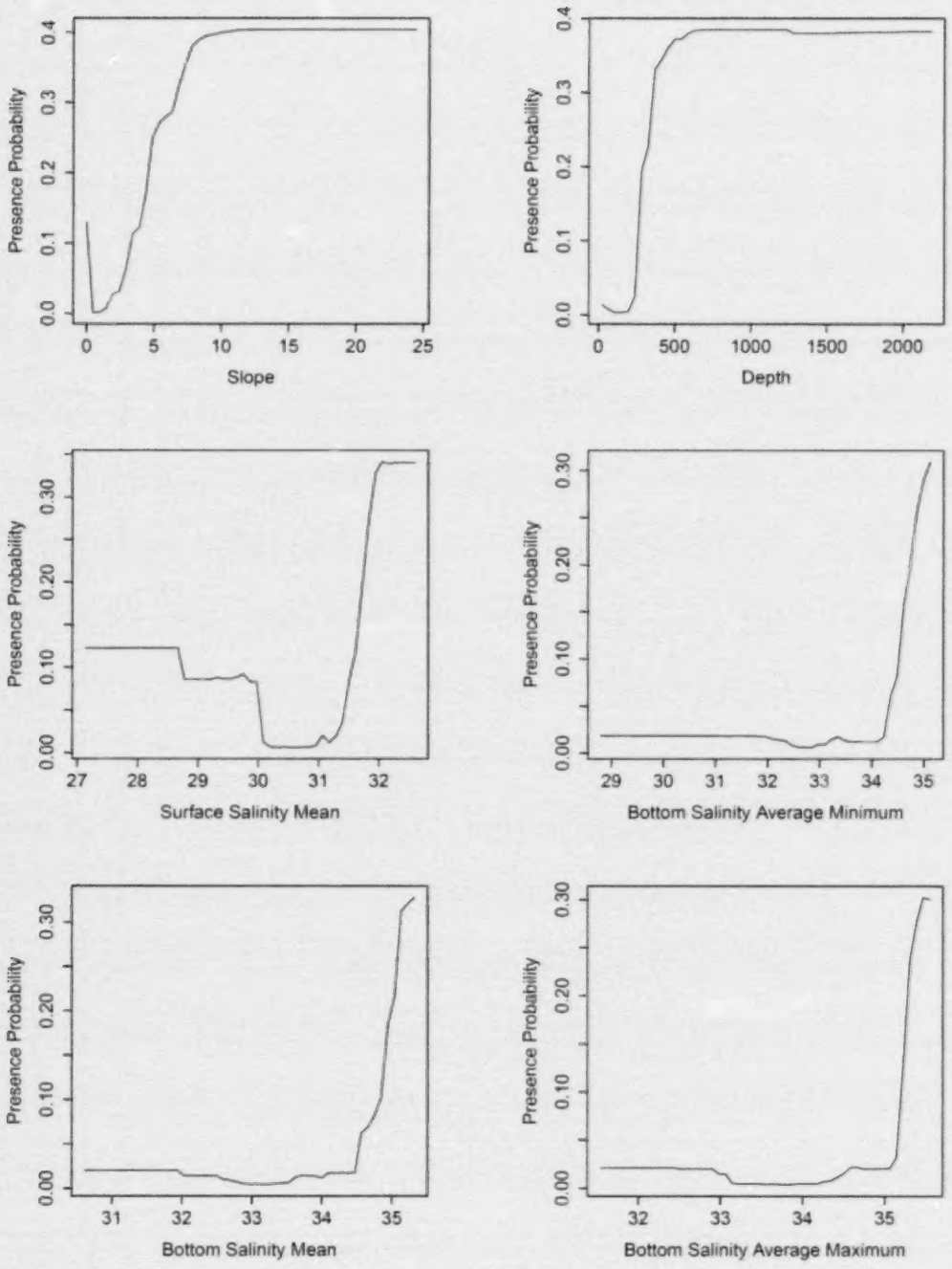

Figure 85. Partial dependence plots of the top six predictors from the random forest model of large gorgonian unbalanced presence and absence data collected from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1967 and 2015, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all available large gorgonian coral records and an unbalanced species prevalence (Model 3) was selected as the best predictor of large gorgonian coral distribution in the Maritimes Region (Figure 80). Model 1 (balanced species prevalence) was considered a poor predictor of presence probability of large gorgonians due to its exaggeration of high presence probability on the banks of the eastern Scotian Shelf. This phenomenon is likely due to random down-sampling of the absence data. Model 2, which was generated using the same presence-absence dataset but using all absence data, produced a much more realistic presence probability surface with less exaggeration beyond the location of presence points. The additional presence records added to Model 3 produced the highest AUC and sensitivity and specificity measures of all three models. Although the presence probability surface was similar to that of Model 2, this model predicted higher large gorgonian presence probability along the eastern Scotian Shelf slope and in its canyons, providing a more accurate depiction of the distribution of large gorgonians in the region based known locations.

Validation of Selected Model Using Independent Data

Figure 86 shows the predicted presence probabilities of large gorgonians generated from Model 3 at the location of large gorgonian records from two independent data sources. There is relatively good congruence between the location of large gorgonian records from these independent data sources and areas of presence predicted by the model. Many of the NOAA records were concentrated in the Northeast Channel and along the eastern slope where there is a high predicted presence probability of large gorgonians (top map in Figure 86). Of the 168 large gorgonian records 27% were predicted as absence based on the prevalence threshold of 0.09 (yellow symbols in figure). These were located mainly in the eastern Gulf of Maine and Bay of Fundy area. Several records occurred in deeper waters off the shelf in an area considered as extrapolated by the model.

A similar pattern is seen in the large gorgonian records from the Gass (2002) and Breeze et al. (1997) reports, with many records concentrated in areas of high predicted presence probability from the model. However, about 40% of the large gorgonian records from this source were predicted as absence based on the prevalence threshold. These records were located mainly in the shallow waters east and southeast of Cape Breton, on Misaine and Banquereau Banks, and in Emerald Basin.

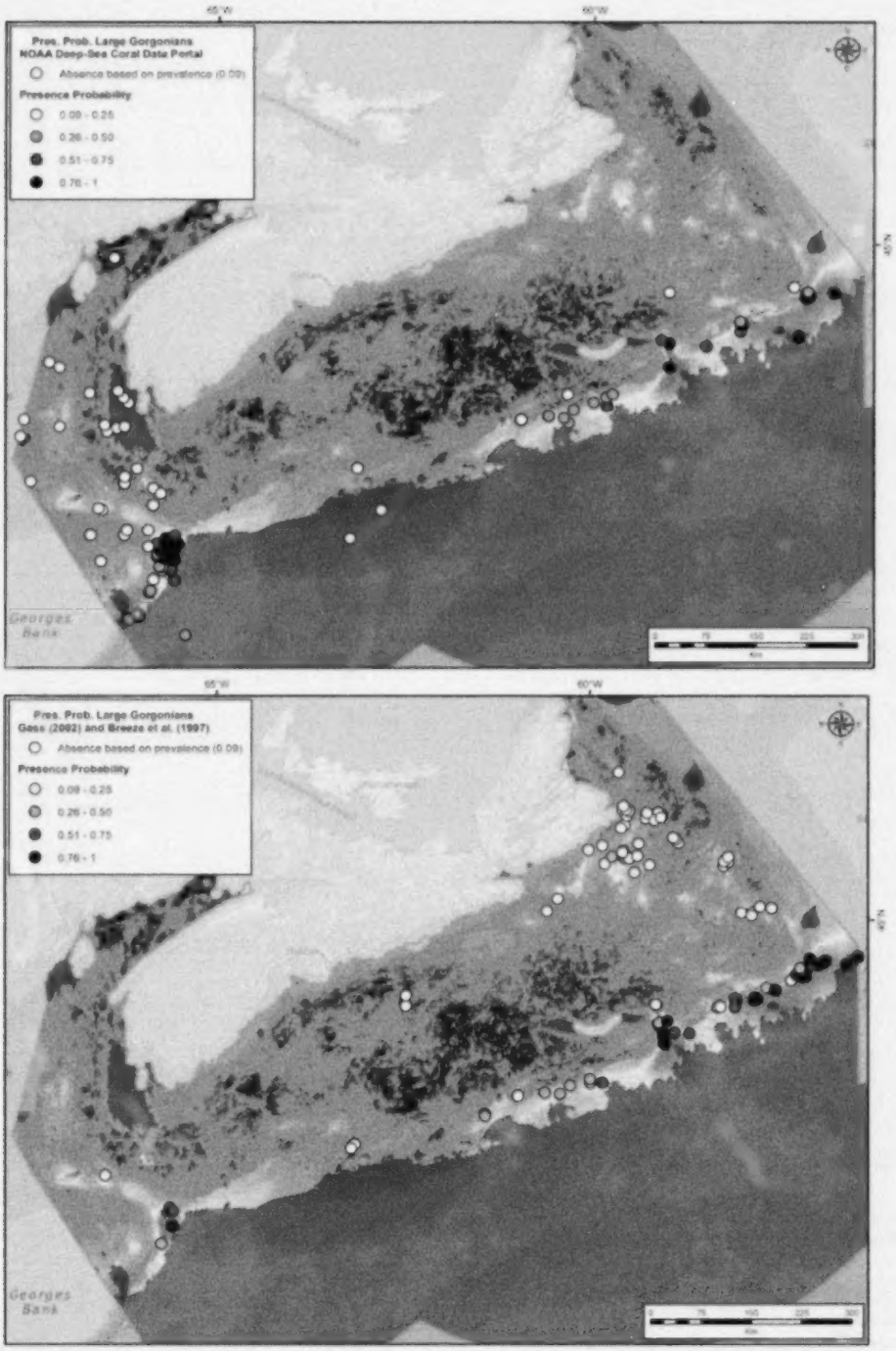

Figure 86. Validation of large gorgonian coral presence probability from Model 3 using independent data. Presence probability values were extracted to the location of large gorgonian coral records from the NOAA Deep-Sea Coral Data Portal (top figure) and from the Gass (2002) and Breeze et al. (1997) reports.

Prediction of Large Gorgonian Coral Biomass Using Random Forest

The accuracy measures of the regression random forest model on mean large gorgonian coral biomass per grid cell from DFO multispecies trawl surveys are presented in Table 24. The highest R^2 value was 0.975, while the average was 0.285 ± 0.410 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.016 ± 0.016 SD (Table 24). The standard deviation was as high as the mean indicating high variability between model folds. This model explained a relatively high percentage of variance in the biomass data (average = 24.53% ± 7.08 SD).

Figures 87 and 88 show the predicted biomass surface of large gorgonians. The majority of the spatial extent was predicted to have low (0 – 2.19 kg) large gorgonian biomass. The slope between Haldimand Canyon and Stone Fence had the highest predicted biomass up to 34.72 kg. This area of high biomass was associated with a cluster of large biomass values (Figure 88). Several canyons that intersect the eastern Scotian Slope, such as The Gully and Shortland Canyon, and the Northeast Channel on the western Scotian Slope, were predicted to have a moderate to high biomass.

Table 24. Accuracy measures from 10-fold cross validation of a random forest model of average large gorgonian coral biomass (kg) per grid cell recorded from DFO multispecies trawl surveys in the Maritimes Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	0.764	0.375	0.007	27.53
2	0.005	0.802	0.015	30.86
3	9.702×10^{-5}	0.150	0.003	24.62
4	0.002	1.752	0.032	37.40
5	0.001	0.127	0.002	24.18
6	0.251	1.461	0.027	19.98
7	8.836×10^{-5}	0.152	0.003	23.16
8	0.975	2.597	0.048	12.11
9	0.856	1.196	0.022	17.57
10	1.540×10^{-6}	0.102	0.002	27.90
Mean	**0.285**	**0.872**	**0.016**	**24.53**
SD	**0.410**	**0.859**	**0.016**	**7.08**

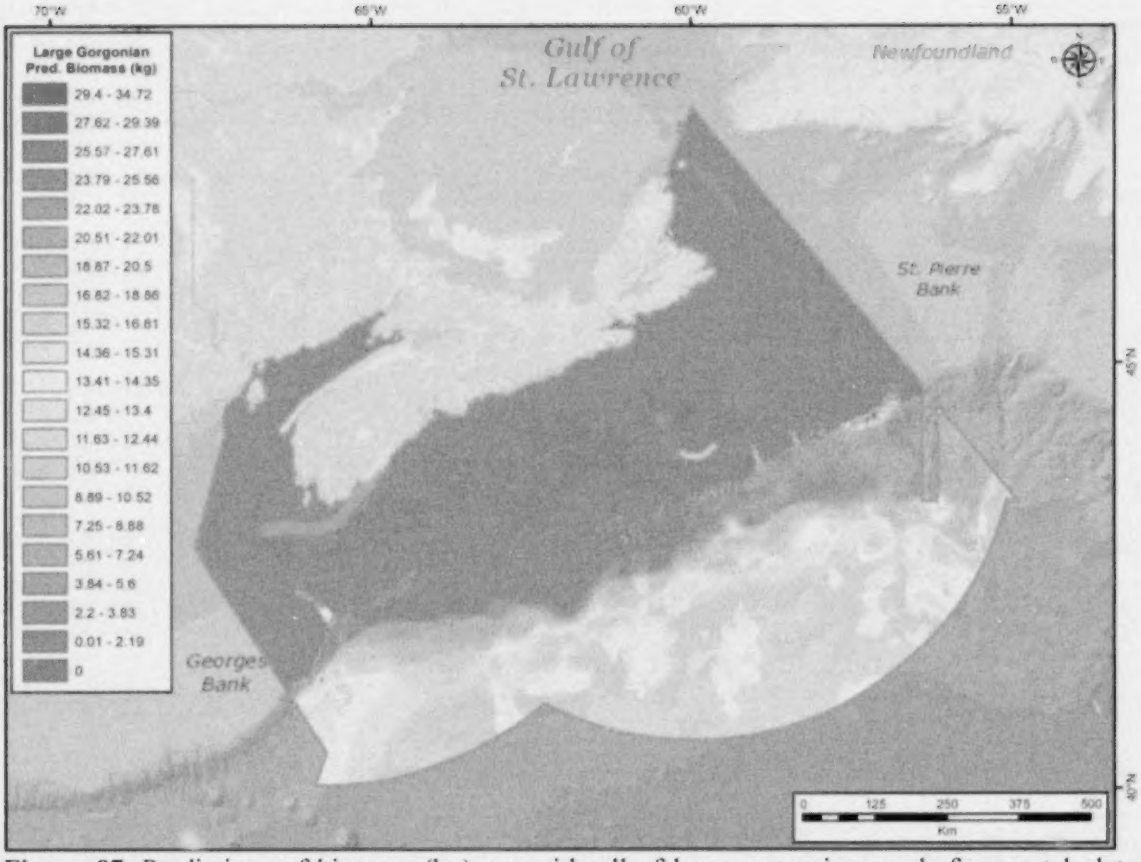

Figure 87. Predictions of biomass (kg) per grid cell of large gorgonian corals from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2002 and 2015.

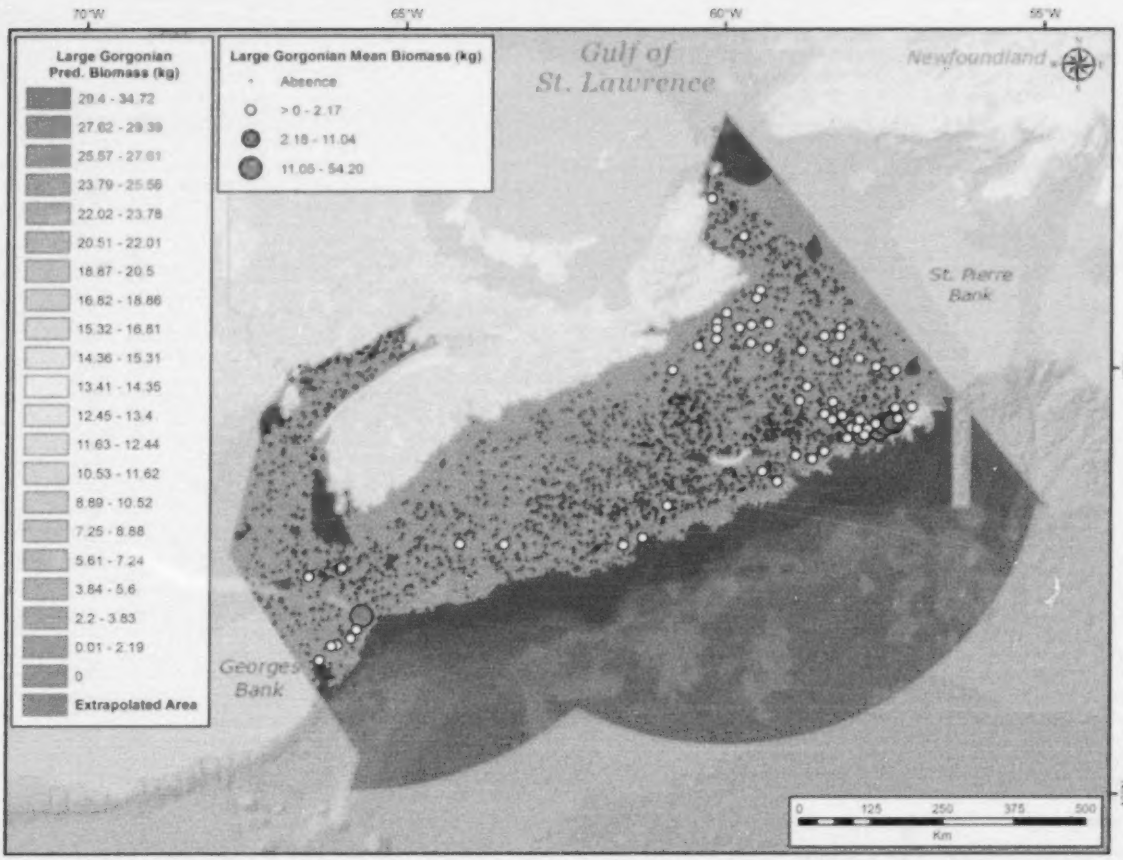

Figure 88. Predictions of biomass (kg) per grid cell of large gorgonian corals from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2002 and 2015. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting large gorgonian coral biomass are shown in Figure 89. Like the presence-absence models, Slope was the most important variable for predicting the biomass of large gorgonian corals, followed very distantly by Bottom Salinity Mean and the other variables in the model. The partial dependence of large gorgonian biomass on the top 6 most important variables is shown in Figure 90. Predicted biomass was highest at slopes greater than 11° and bottom salinity values greater than 35.

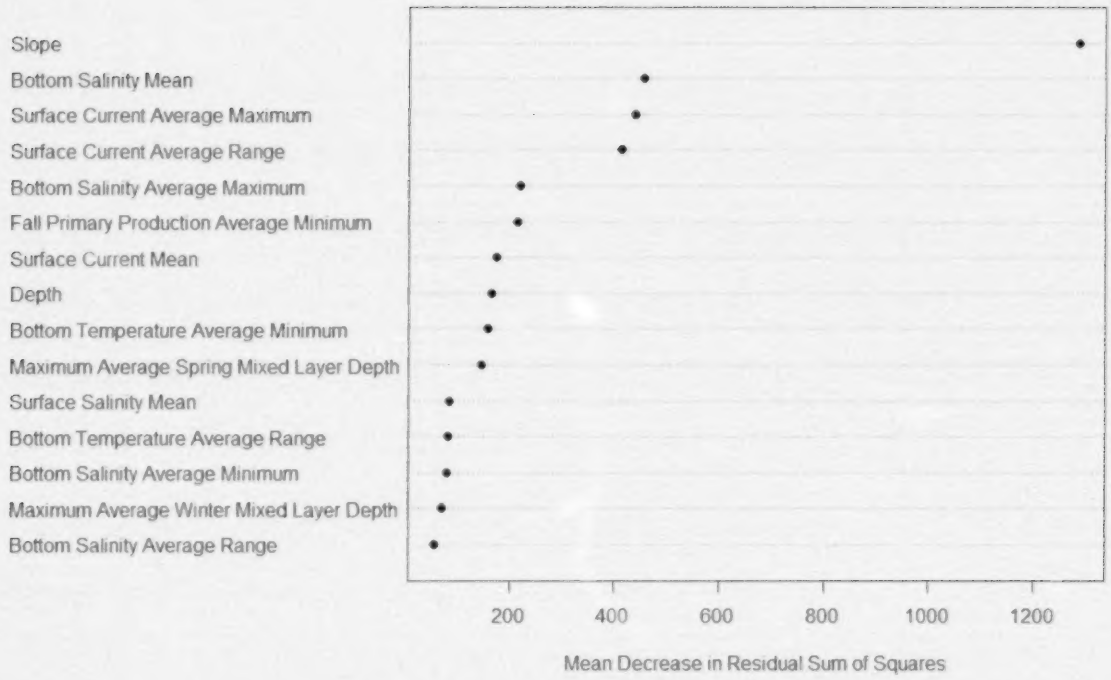

Figure 89. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on large gorgonian coral mean biomass averaged per grid cell. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

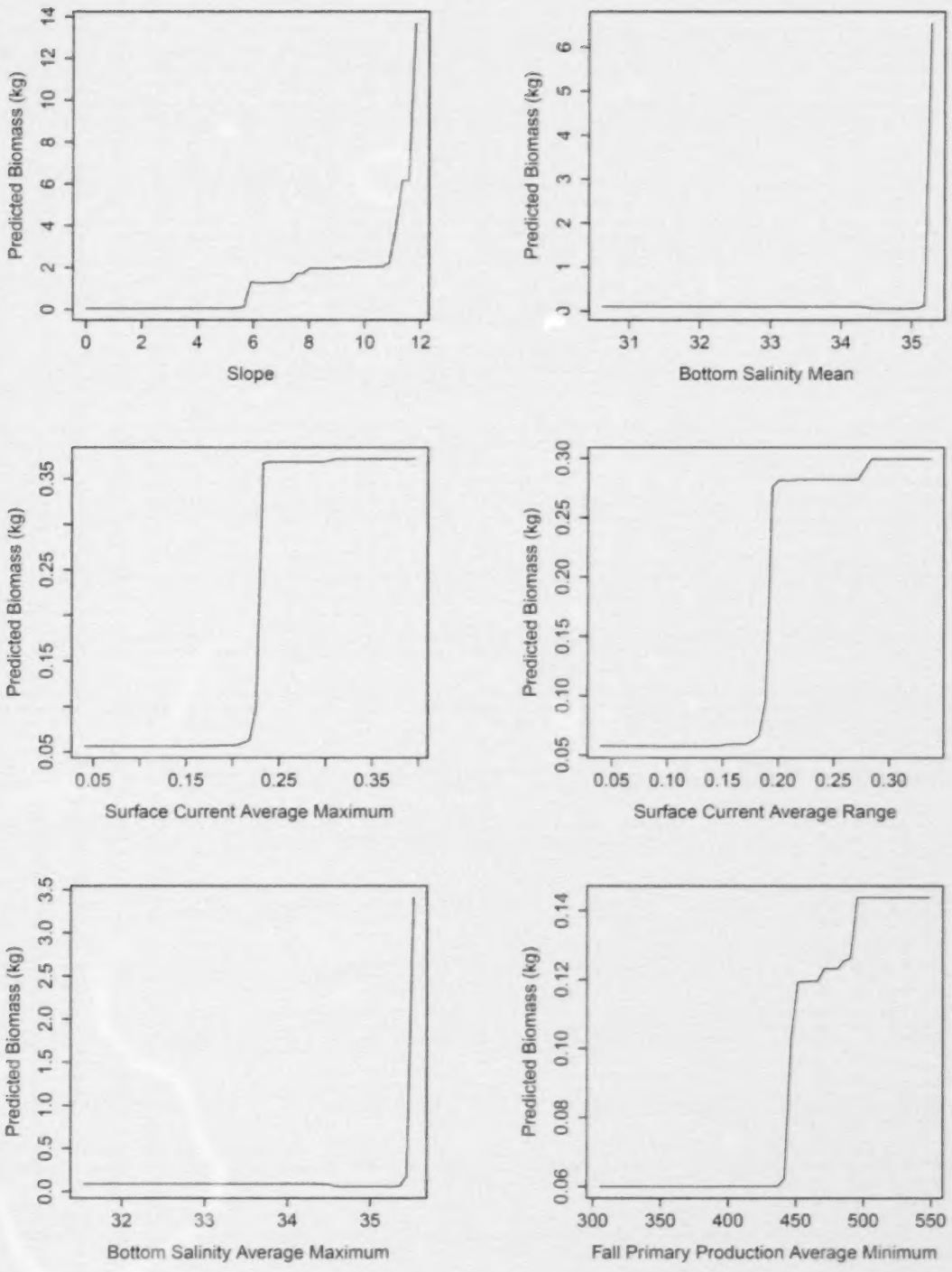
Predicted Biomass (kg)
0 2 4 6 8 10 12 14
0 2 4 6 8 10 12
Slope
Predicted Biomass (kg)
0 1 2 3 4 5 6
31 32 33 34 35
Bottom Salinity Mean
Predicted Biomass (kg)
0.05 0.15 0.25 0.35
0.05 0.15 0.25 0.35
Surface Current Average Maximum
Predicted Biomass (kg)
0.05 0.10 0.15 0.20 0.25 0.30
0.05 0.10 0.15 0.20 0.25 0.30
Surface Current Average Range
Predicted Biomass (kg)
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
32 33 34 35
Bottom Salinity Average Maximum
Predicted Biomass (kg)
0.06 0.08 0.10 0.12 0.14
300 350 400 450 500 550
Fall Primary Production Average Minimum

Figure 90. Partial dependence plots of the top six predictors from the random forest model of large gorgonian coral biomass data collected within the Maritimes Region, ordered left to right from the top. Predicted biomass (kg) is shown on the *y*-axis.

Small Gorgonian Corals

Data Sources and Distribution

Figure 91 shows the distribution of available small gorgonian records in the Maritimes Region. There was relatively good congruence in the spatial distribution of records originating from the different data sources. DFO multispecies trawl survey records were concentrated on several shallow banks on the eastern Scotian Shelf and slope, and in the Laurentian Channel. The scientific survey and NOAA were concentrated along the slopes, and in the Northeast Channel and the deeper waters beyond.

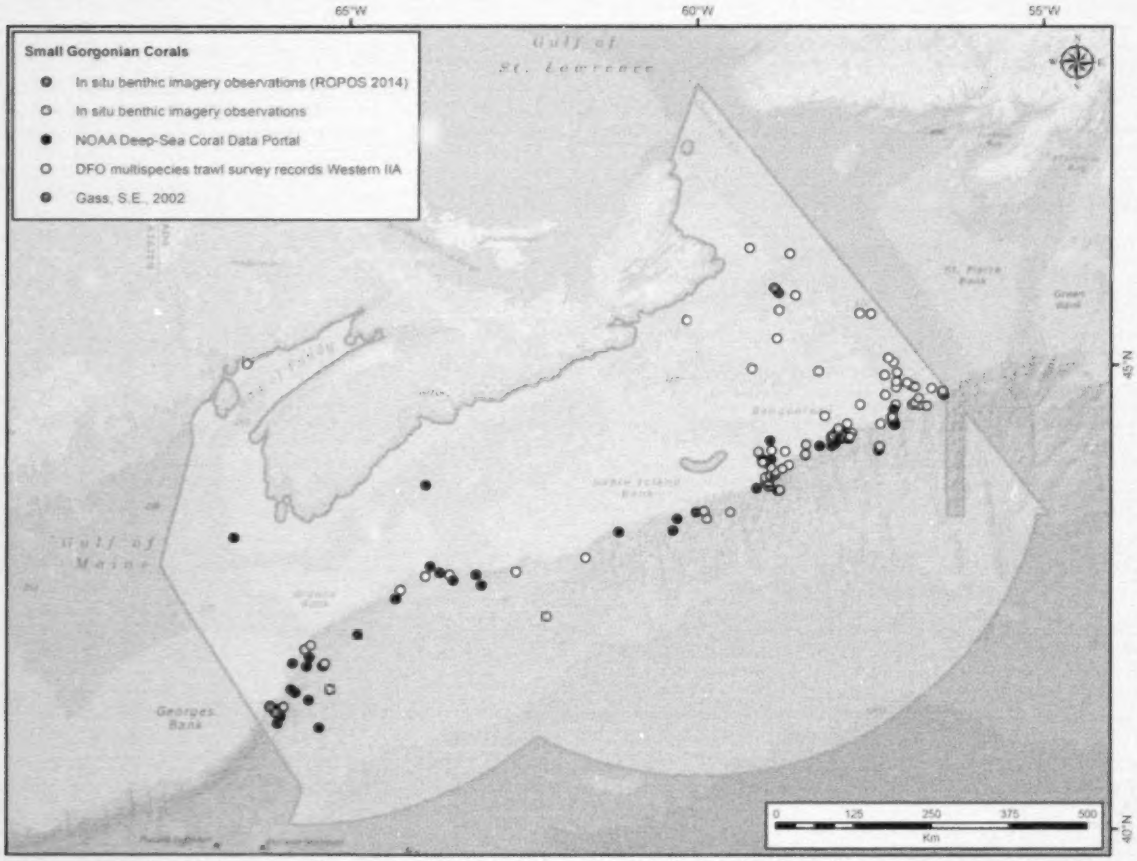

Figure 91. Available small gorgonian coral presence data in the Maritimes Region from Gass (2002), scientific missions, the NOAA Deep-Sea Coral Data Portal, and DFO multispecies research vessel surveys.

Initial random forest models of small gorgonians were run using only catch data originating from DFO multispecies trawl surveys (Western IIA gear). This data was collected over a period of 8 years from 2002 to 2014 (Table 25). This dataset was highly imbalanced, consisting of 36 presence and 1817 absence records (Figure 92). Absence records were distributed relatively evenly across the Scotian Shelf and Bay of Fundy. The highest mean biomass record (up to 0.34 kg) occurred on the slope between Emerald and LaHave Banks, while a cluster of mid-range biomass values occurred in the southern Laurentian Channel. A single catch of small gorgonian corals was recorded in the Bay of Fundy south of New Brunswick.

Table 25. Number of presence and absence records of small gorgonian catch recorded from DFO multispecies trawl surveys conducted between 2002 and 2014 in the Maritimes Region.

Year	Total number of presences	Total number of absences
2002	6	324
2003	5	319
2005	6	206
2006	3	248
2007	7	75
2009	6	191
2011	2	257
2014	1	197

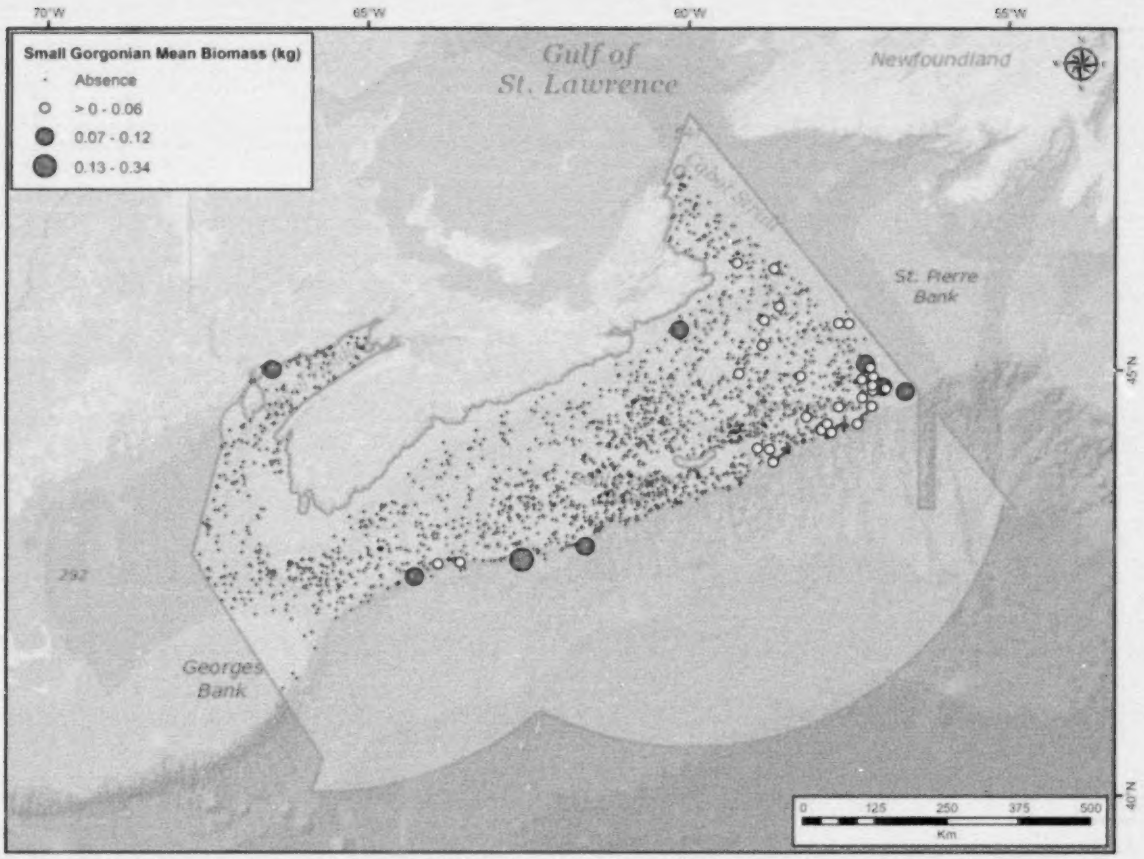

Figure 92. Mean biomass (kg) per grid cell of small gorgonian coral catch recorded from DFO multispecies trawl surveys from 2002 to 2014 within the Maritimes Region. Also shown are absence records from the same surveys.

Model 1 – Balanced Species Prevalence

Accuracy measures (mean AUC, sensitivity, and specificity) for the random forest model on balanced species prevalence (36 presences and 36 absences; Model 1) are presented in Table 26. The highest mean AUC of 0.922 was associated with Model Run 6 and is therefore considered the optimal model for the prediction of the small gorgonian coral response data. The sensitivity and specificity measures of this model run were 0.722 and 0.806, respectively. The confusion matrix of the optimal model is also presented in Table 2. Class error for the presence class was somewhat moderate (0.278).

Table 26. *Accuracy measures for all 10 model repetitions of 10-fold cross validation of a random forest model of presence and absence of small gorgonian corals within the Maritimes Region. The confusion matrix is shown for the model with the highest AUC value (Model Run 6) which is considered the optimal model for predicting the presence probability of small gorgonian corals in the region.*

Model Run	AUC	Sensitivity	Specificity
1	0.848	0.722	0.750
2	0.822	0.778	0.722
3	0.855	0.667	0.667
4	0.846	0.639	0.667
5	0.883	0.667	0.778
6	**0.922**	**0.722**	**0.806**
7	0.773	0.611	0.722
8	0.788	0.722	0.722
9	0.810	0.667	0.750
10	0.905	0.750	0.806
Mean	**0.845**	**0.694**	**0.739**
SD	**0.049**	**0.052**	**0.049**

Confusion matrix of model with highest AUC:

Observations	Predictions		Total n	Class error
	Absence	Presence		
Absence	29	7	36	0.194
Presence	10	26	36	0.278

The presence probability prediction surface of small gorgonian corals is presented in Figure 93. The highest predictions of presence probability occurred in southwestern Laurentian Channel off Cape Breton. Areas of higher presence probability corresponded well with the spatial distribution of presence records (see Figure 94), however, the model appears to greatly extrapolate high areas of presence probability beyond these locations.

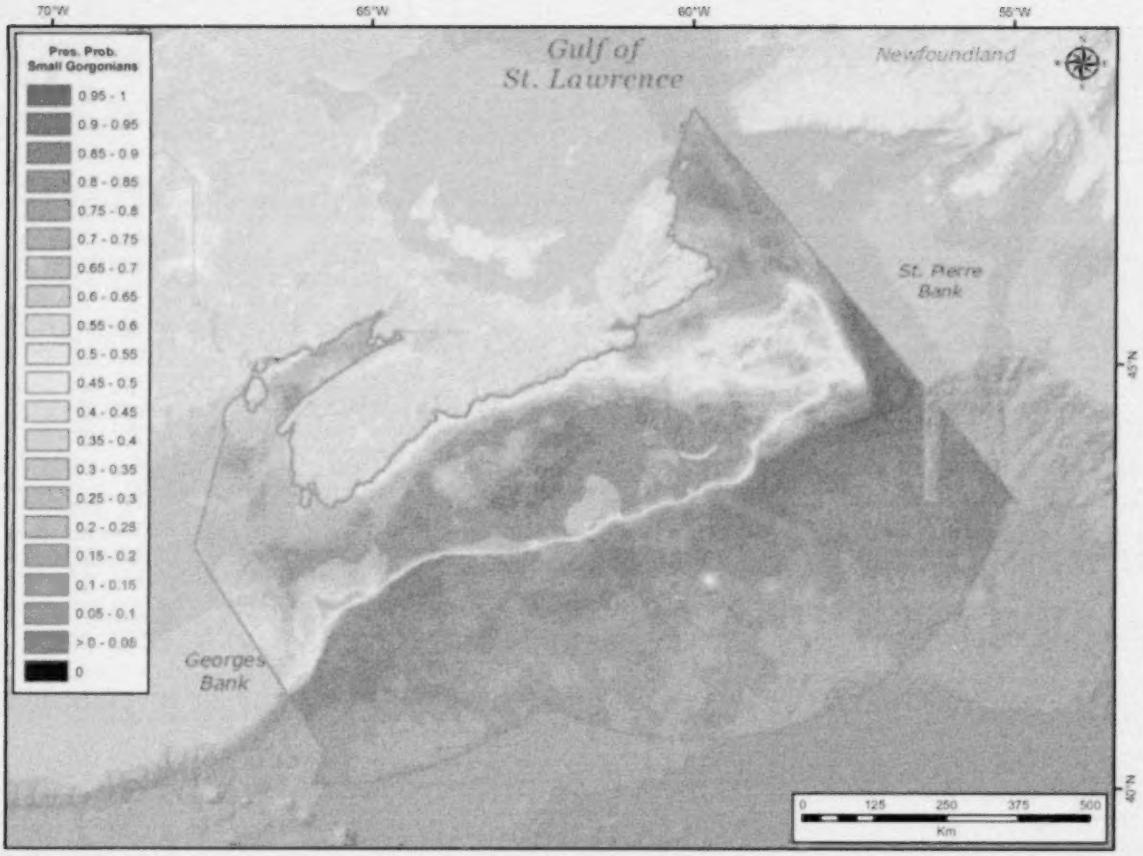

Figure 93. Predictions of presence probability (Pres. Prob.) from the optimal random forest model of small gorgonian coral presence and absence data collected from DFO multispecies trawl surveys in the Maritimes Region between 2002 and 2014.

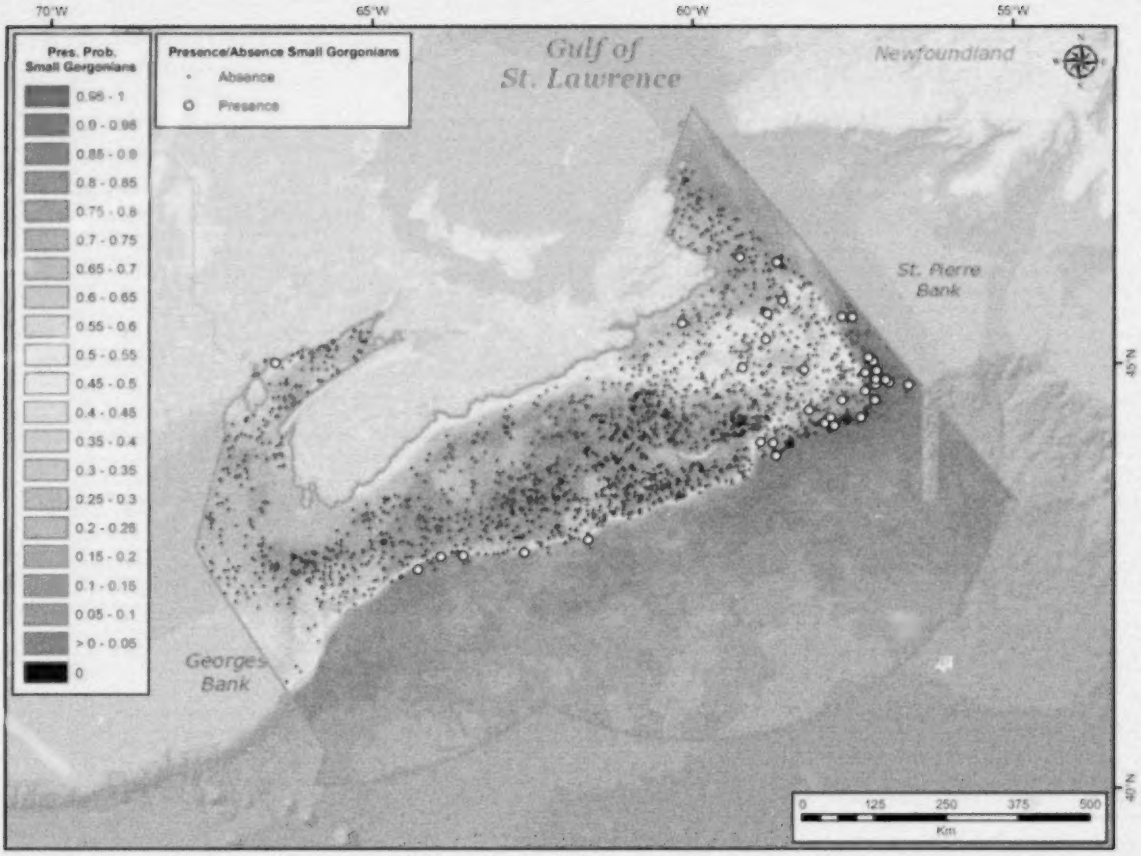

Figure 94. Presence and absence observations and predictions of presence probability (Pres. Prob.) of the optimal random forest model of small gorgonian presence and absence data recorded from DFO multispecies trawl surveys in the Maritimes Region between 2002 and 2014.

The actual presence and absence data observations (36 presences and 36 absences) used in the optimal run of Model 1 showed some slight spatial bias across the study area (Figure 95). Despite the occurrence of absence records off Cape Breton, none were selected from this area during the random down-sampling of the data prior to modelling, likely causing the moderate to high predictions of presence probability there. Also shown in this figure are the areas of model extrapolation. Deep water beyond the Scotian Shelf is considered extrapolated area, as well as smaller areas off southwestern Nova Scotia and northeast tip of Cape Breton. Much of the western and central portion of the study area is considered extrapolated area, as well as a large area off the coast of Cape Breton and in the northern Laurentian Channel where there were pockets of high presence probability of small gorgonians.

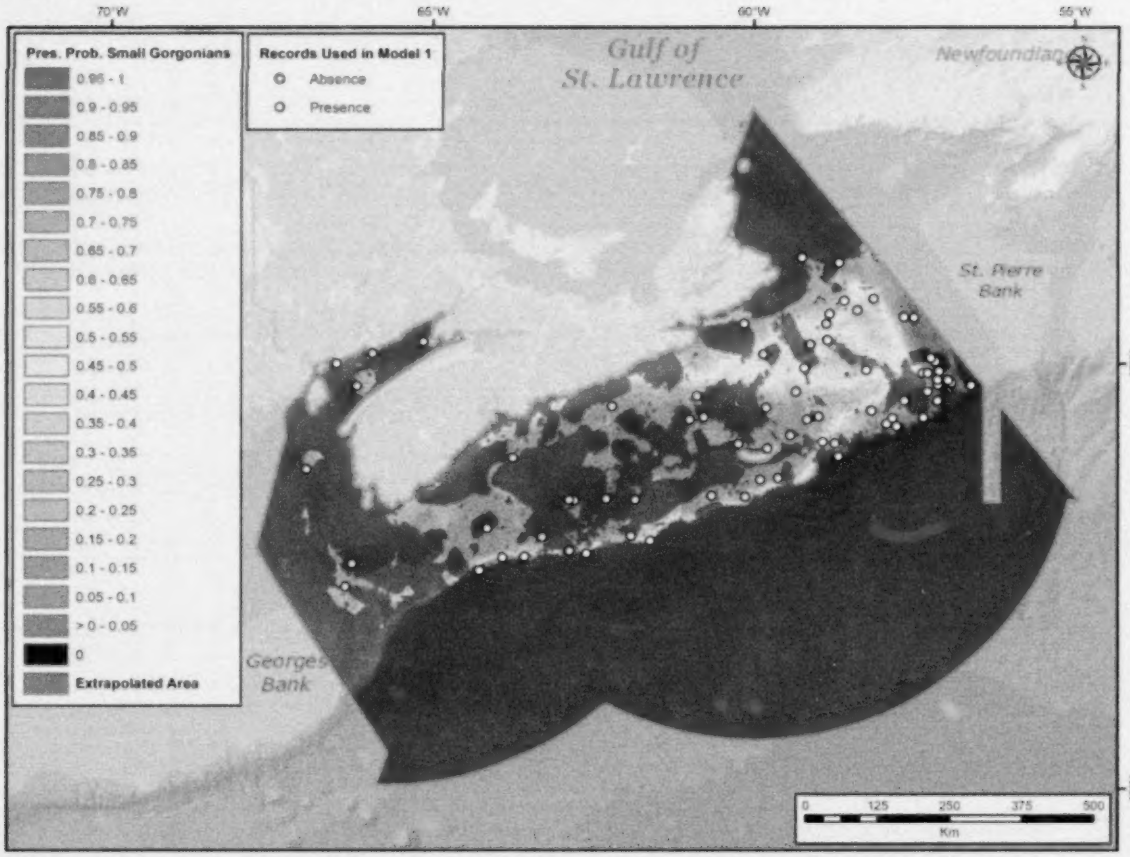

Figure 95. Map of the 72 data observations (36 presences and 36 absences) of small gorgonian corals used in the optimal random forest Model 1 on balanced species prevalence. Also shown is the predicted presence probability (Pres. Prob.) of small gorgonian corals generated from Model 1 and areas of model extrapolation.

Of all 66 environmental predictor variables used in the model, Maximum Average Spring Mixed Layer Depth was the most important for the classification of the small gorgonian presence and absence data (Figure 96). Prior to spatial interpolation, this variable displayed a right-skewed distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed a spatial pattern to data points over- and under-predicted by a normal distribution with over-predicted points located along the coast of Nova Scotia and in the deepest regions of the study extent, and under-predicted points located across the centre of the study extent and off southwestern Nova Scotia. This variable was followed more distantly in terms of its Mean Decrease in Gini Value by Depth and Bottom Salinity Average Range. Surface current and bottom temperature variables had high importance in this model. The partial dependence plots for the top 6 most important predictors are shown in Figure 97. Presence probability of small gorgonians was highest at Maximum

Average Spring Mixed Layer Depth values of 20 m and greater. Values in this range coincided with both over- and under-predicted data points mainly in the deeper portion of the study extent but also in Gulf of Maine and the southern Laurentian Channel. The fit between observed and predicted values in the kriging model was fair, with slight over-prediction of Maximum Average Spring Mixed Layer Depth values at 20 m and greater, which would still coincide with the range of high presence probabilities indicated in the partial plot. Along the Depth gradient, presence probability of small gorgonian corals increased and remained high at 200 m.

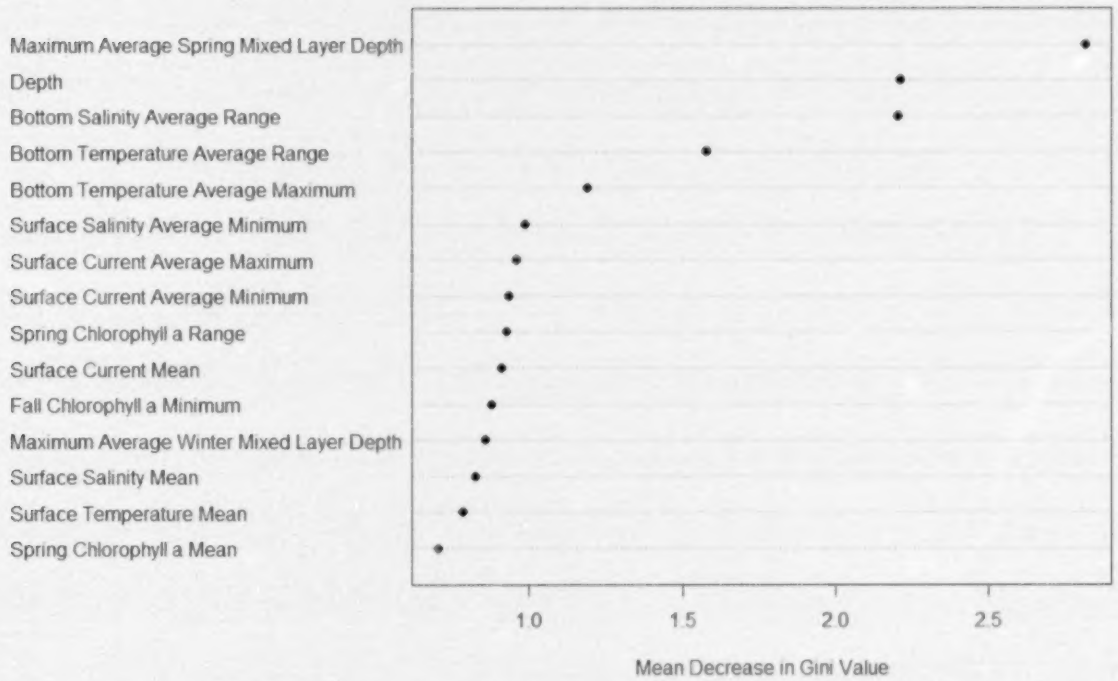

Figure 96. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the optimal random forest model predicting small gorgonian coral presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

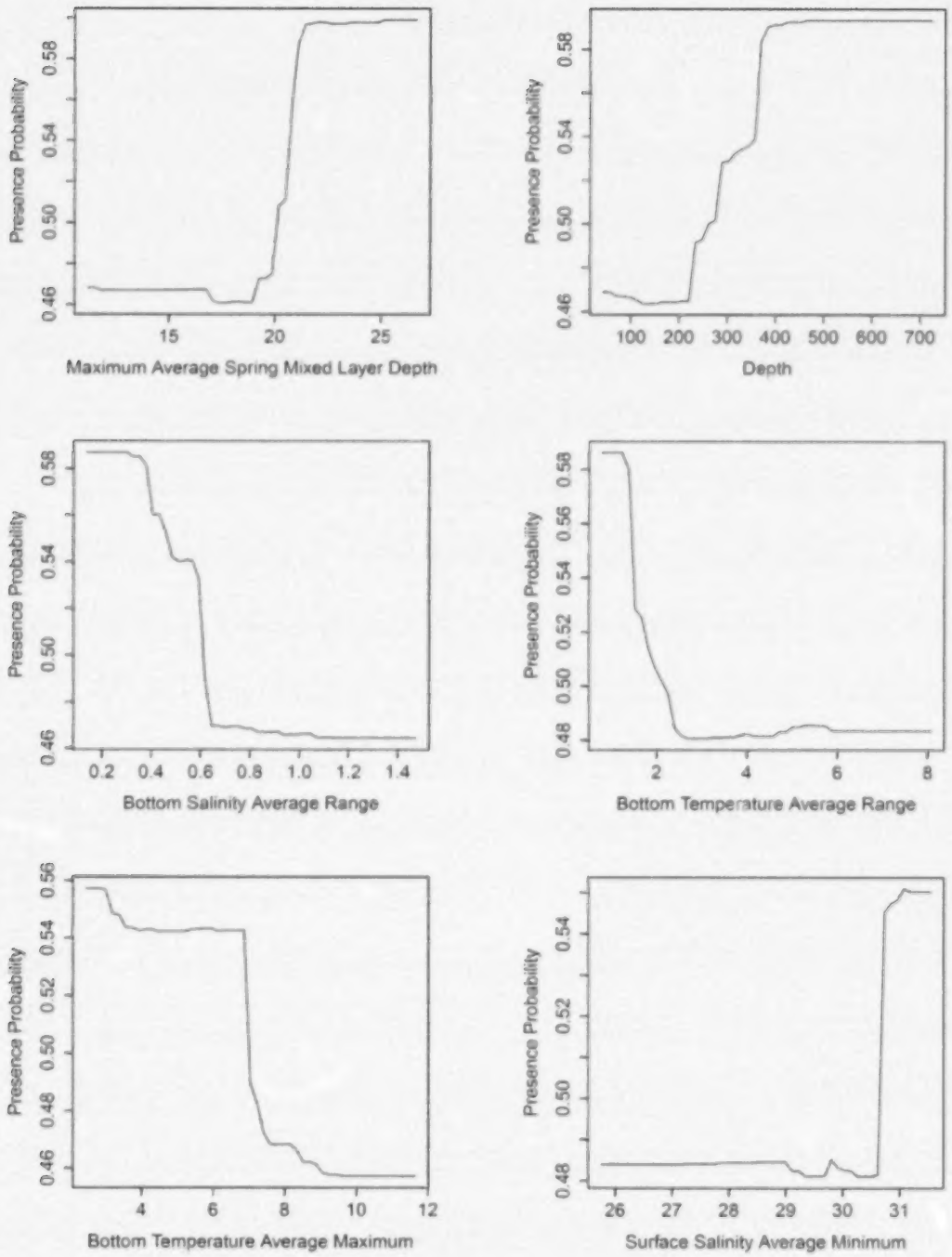

Figure 97. Partial dependence plots of the top six predictors from the optimal random forest model of small gorgonian presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 2 – Unbalanced Data and Threshold Equal to Species Prevalence

Table 27 shows the accuracy measures for the random forest model on all small gorgonian coral presence and absence data (36 presences and 1817 absences; Model 2) and a threshold equal to species prevalence (0.02). The average AUC calculated from Model 2 was 0.797, lower than that of Model 1. Class error of the absence class was similar to that of Model 1 (0.183 compared to 0.194 from Model 1), however class error for the presence class was slightly higher (0.389 compared to 0.278 from Model 1). Sensitivity was relatively low for Model 2 (0.611) but similar to the average sensitivity from Model 1 (0.694). Specificity of Model 2 was higher than that of Model 1 (0.817 versus 0.739).

The predicted presence probability surface of small gorgonian corals generated from Model 2 is shown in Figure 98. The southwestern Laurentian Channel is the only area predicted to have a relatively high presence probability of small gorgonian corals, likely due to the concentrations of presence records that reside there (Figure 99). The remainder of the Laurentian Channel was predicted to have low to moderate probability of presence of small gorgonian corals. Much of central Scotian Shelf and Bay of Fundy is predicted to have zero or low presence probability of small gorgonians.

Table 27. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of small gorgonians within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.930		Absence	Presence				
2	0.554	Absence	1484	333	1817	0.183	0.611	0.817
3	0.626	Presence	14	22	36	0.389		
4	1.000							
5	0.788							
6	0.924							
7	0.515							
8	0.995							
9	0.996							
10	0.640							
Mean	**0.797**							
SD	**0.196**							

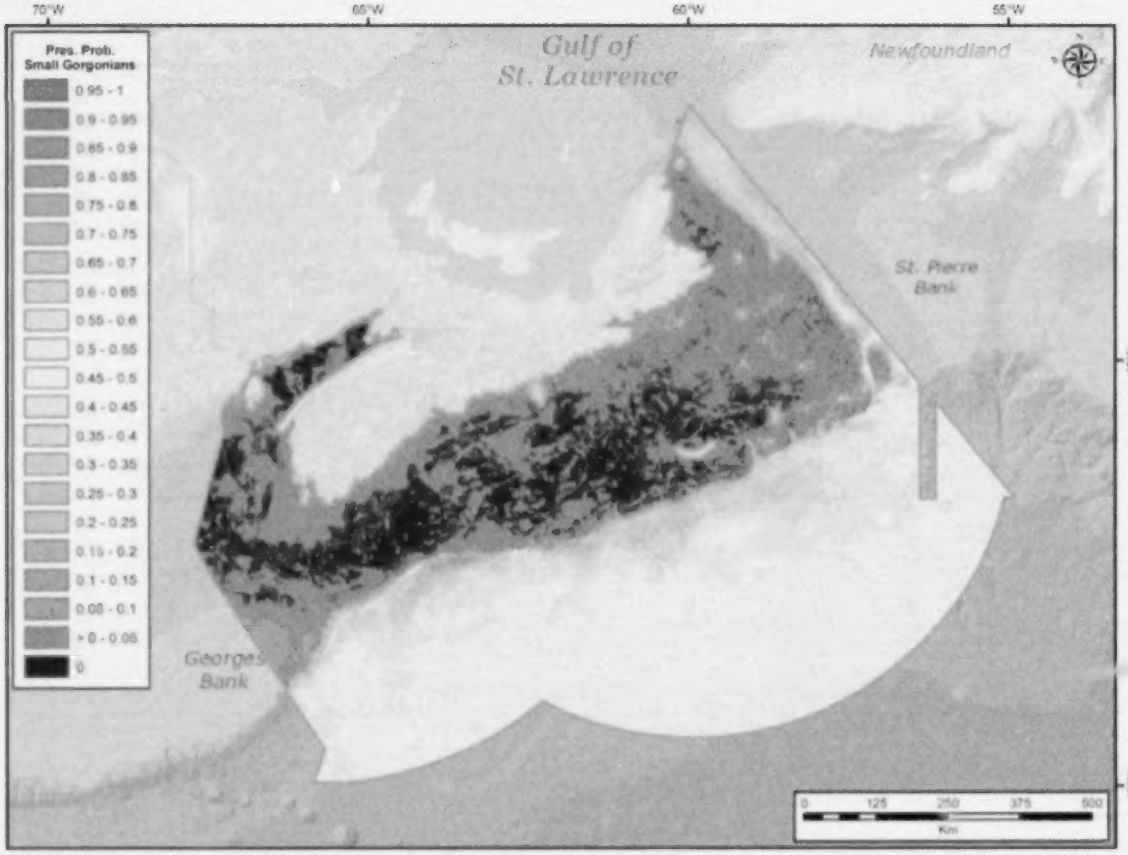

Figure 98. Predictions of presence probability (Pres. Prob.) of small gorgonian corals based on a random forest model on unbalanced presence and absence small gorgonian coral catch data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2002 and 2014.

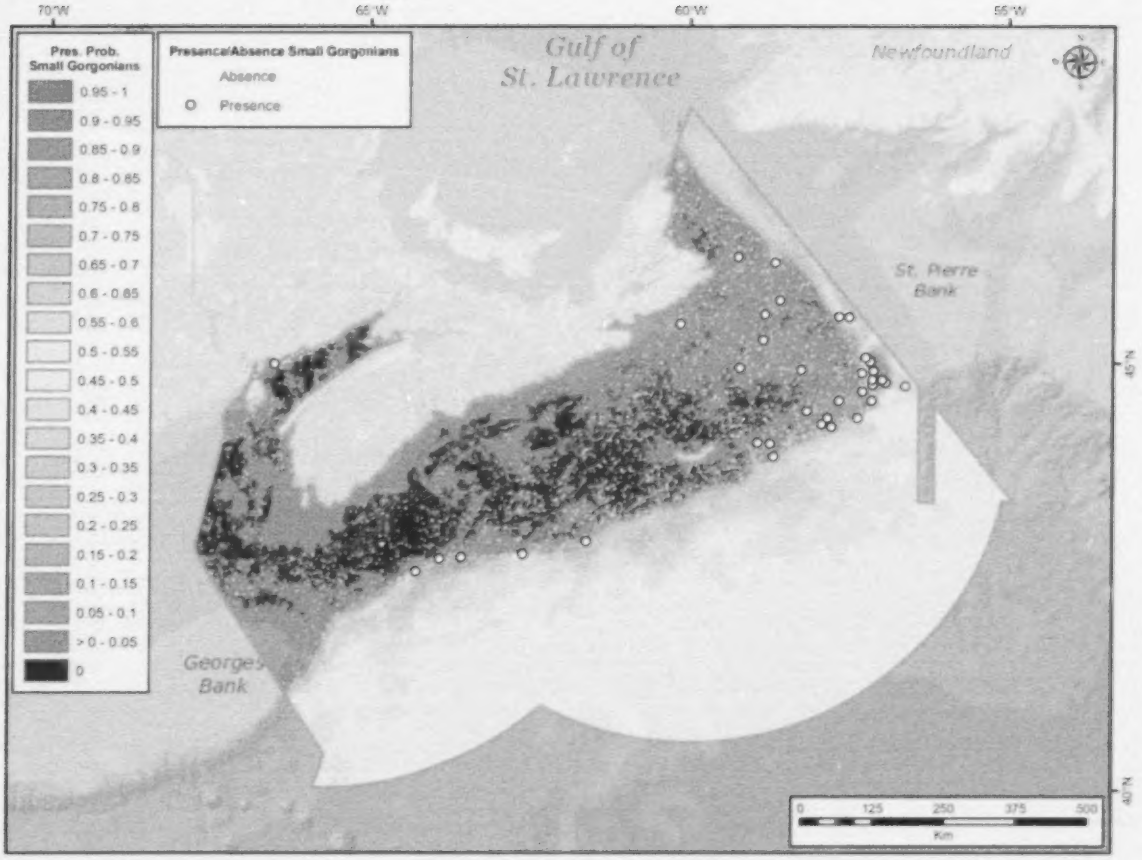

Figure 99. Presence and absence observations and predictions of presence probability (Pres. Prob.) of small gorgonian corals based on a random forest model on unbalanced presence and absence small gorgonian coral catch data collected from DFO multispecies trawl surveys conducted within the Maritimes Region between 2002 and 2014.

The order of importance of the environmental predictor variables in Model 2 was slightly different from that of Model 1 (Figure 100). Depth was the most important variable in Model 2 compared to Maximum Average Spring Mixed Layer Depth in Model 1. Maximum Average Spring Mixed Layer Depth was the 4th most important variable in Model 2. Depth was followed in importance by Bottom Salinity Average Range and Bottom Salinity Average Minimum. Several primary production variables were ranked high in this model. Partial dependence of small gorgonian presence and absence data on the top 6 predictor variables is shown in Figure 101. Presence probability of small gorgonians increased beginning at ~300 m depth and plateaued at ~600 m. Presence probability was highest at low Bottom Salinity Average Range and high Bottom Salinity Average Minimum values.

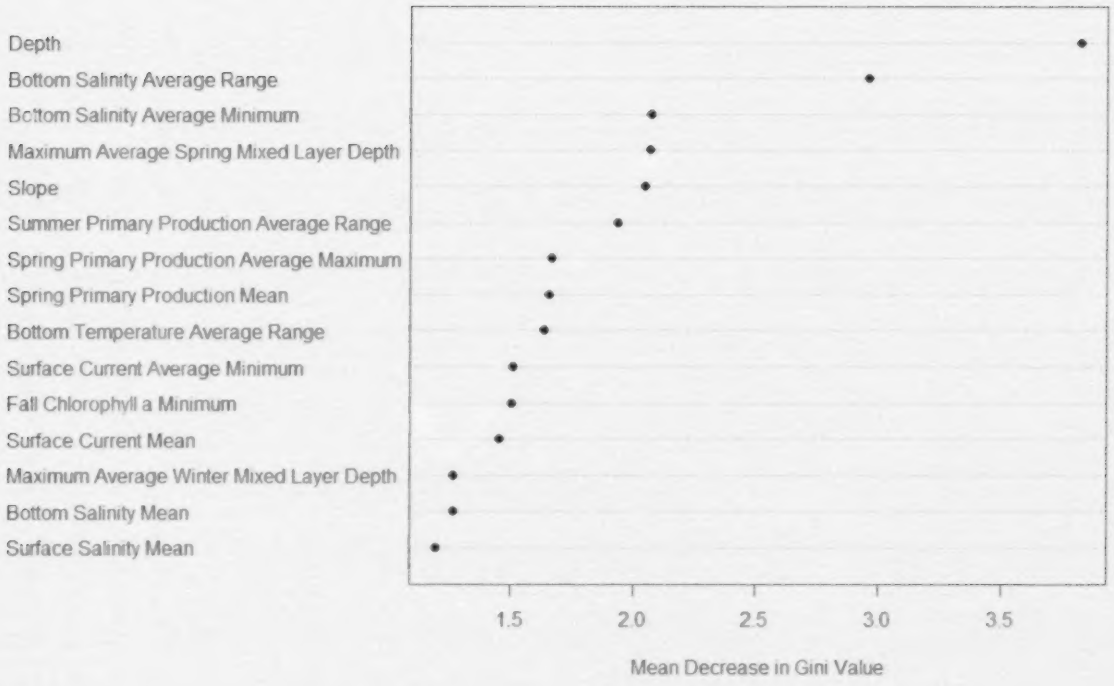

Figure 100. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced small gorgonian coral presence and absence data within the Maritimes Region. The higher the Mean Gini value the more important the variable is for predicting the response data.

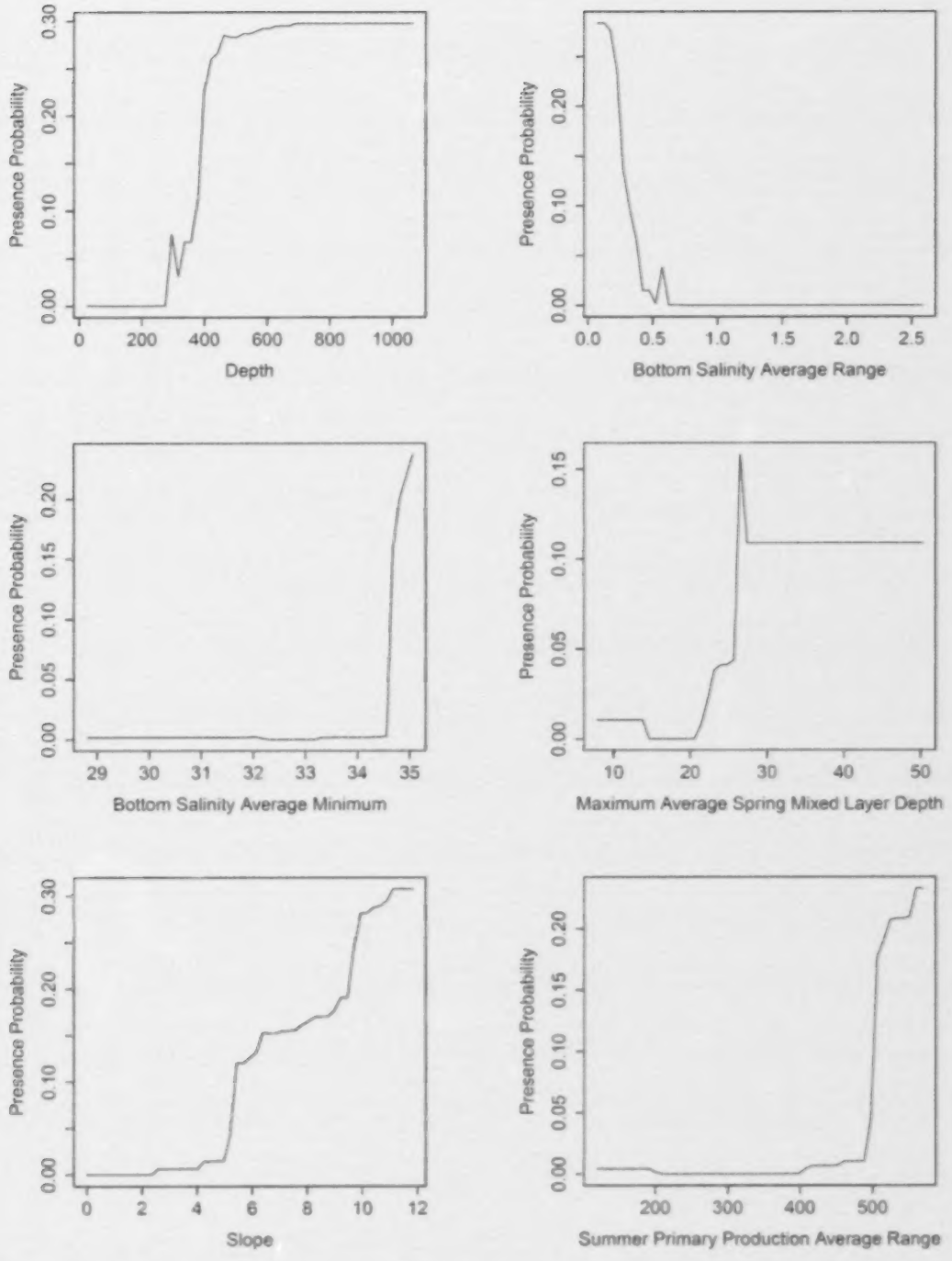

Figure 101. Partial dependence plots of the top six predictors from the random forest model of small gorgonian coral unbalanced presence and absence data collected within the Maritimes Region, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model 3 – Addition of *In Situ* Benthic Imagery Observations

Given the low number of small gorgonian coral presence records in the Maritimes Region, the DFO multispecies trawl survey data were augmented with additional presence records from scientific surveys conducted in the Maritimes Region by DFO and Natural Resources of Canada (NRCan). A total of 85 additional presence records (see Table 28) were added to the DFO multispecies trawl survey dataset after filtering the data so that only one presence record occurred per environmental grid cell. The combined dataset consisting of 121 presence and 1815 absence records was remodelled (termed Model 3) using an unbalanced design and a threshold equal to species prevalence (0.06).

The accuracy measures for random forest Model 3 are shown in Table 29. The average AUC computed from 10-fold cross validation was 0.949 ± 0.033 SD, the highest of all three models. Class error for the presence and absence classes was the lowest of all three models, while sensitivity and specificity were the highest.

The additional presence records expanded the area of high presence probability along the eastern Scotian Slope (Figure 102). The Gully submarine canyon, and the southern Laurentian Channel showed a much greater probability of occurrence of small gorgonian corals compared to Models 1 and 2. The deep waters outside of the Northeast Channel also showed a higher presence probability of small gorgonians compared to the previous models. These areas of higher presence probability corresponded well with the location of the additional presence records from

Table 28. Number of presence and absence records of small gorgonian catch recorded from *in situ* benthic imagery surveys conducted between 1997 and 2014 in the Maritimes Region.

Year	Gear	Total number of presences
1997	Campod	1
2000	NRCan Drop Camera	6
2001	Campod	6
2002	Campod	6
2006	ROPOS	6
2007	ROPOS	29
2008	NRCan Drop Camera	4
2008	Campod	22
2011	Campod	3
2014	Towed Camera	2

Table 29. Accuracy measures and confusion matrix from 10-fold cross validation of a random forest model of presence and absence of small gorgonian corals from DFO multispecies trawl survey records and scientific surveys conducted within the Maritimes Region. Observ. = Observations; Sensit. = Sensitivity, Specif. = Specificity.

Model Fold	AUC	Observ.	Predictions		Total n	Class error	Sensit.	Specif.
1	0.912		**Absence**	**Presence**				
2	0.905	**Absence**	1662	153	1815	0.084	0.876	0.916
3	0.998	**Presence**	15	106	121	0.124		
4	0.946							
5	0.949							
6	0.936							
7	0.988							
8	0.909							
9	0.974							
10	0.971							
Mean	**0.949**							
SD	**0.033**							

DFO and NRCan *in situ* benthic imagery data (Figure 103). The area of extrapolation along the slopes of the Scotian Shelf is reduced with the additional of science survey presence records there (see Figure 104). Figure 105 depicts the classification of small gorgonian presence probability into presence and absence categories based on the prevalence threshold of 0.06. In this map, all presence probability values generated from Model 3 that were greater than 0.06 were classified as presence, while values less than 0.06 were classed as absence. The majority of the Scotian Slope, part of Georges Bank, and the Laurentian Channel were predicted as presence of small gorgonians.

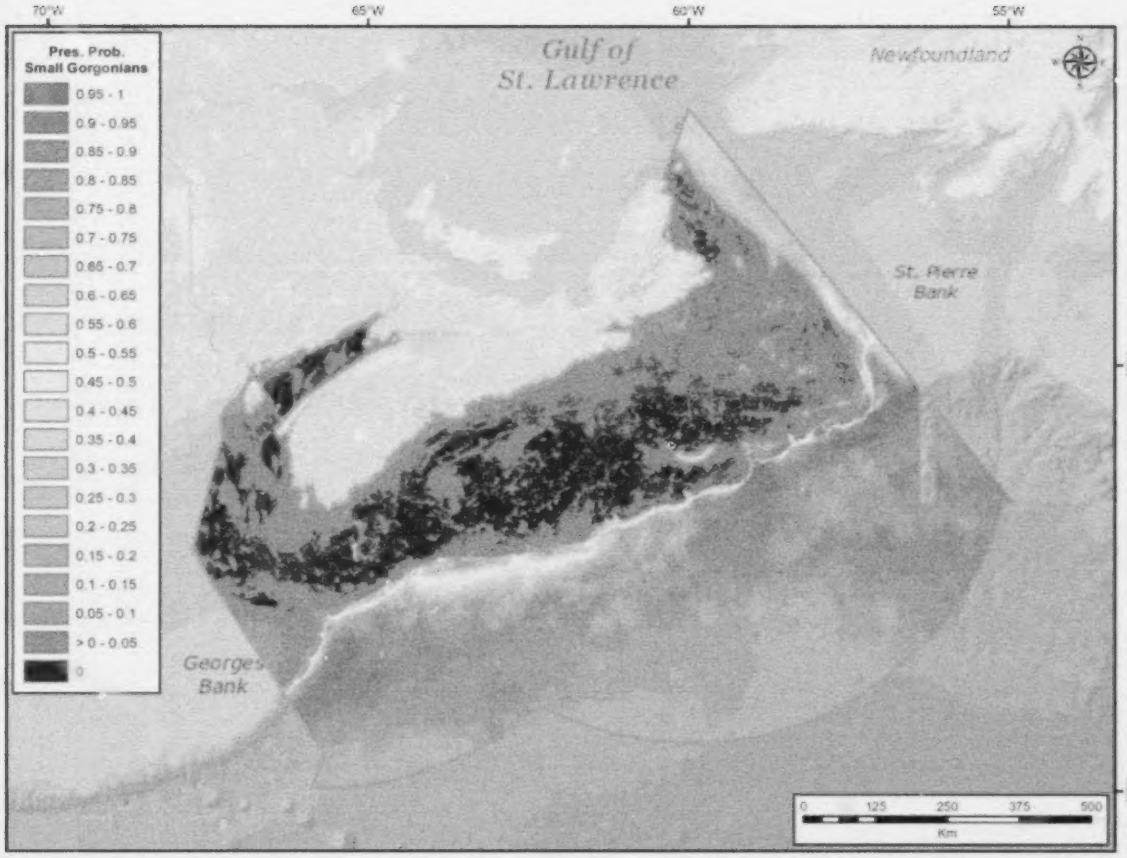

Figure 102. Predictions of presence probability (Pres. Prob.) of small gorgonian corals based on a random forest model on unbalanced presence and absence small gorgonian catch data collected from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1997 and 2014.

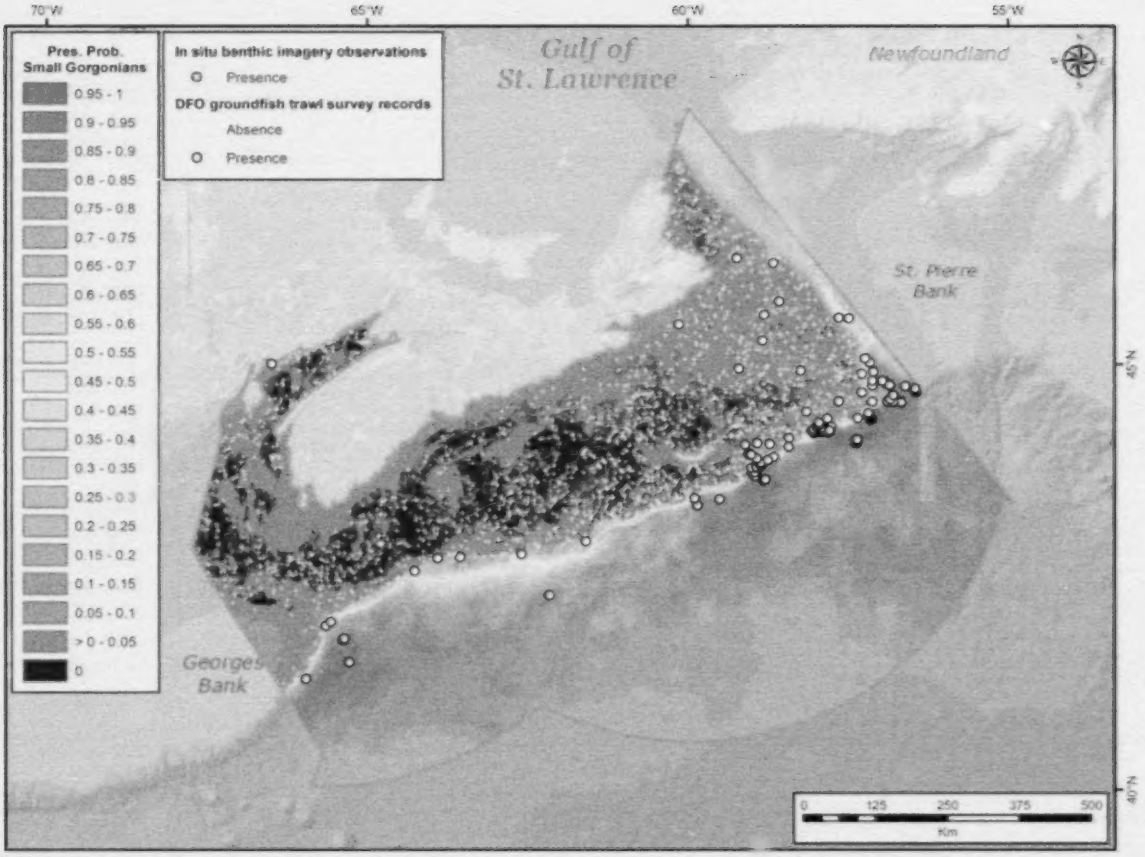

Figure 103. Presence and absence observations and predictions of presence probability (Pres. Prob.) of small gorgonian corals based on a random forest model on unbalanced presence and absence gorgonian catch data collected from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1997 and 2014.

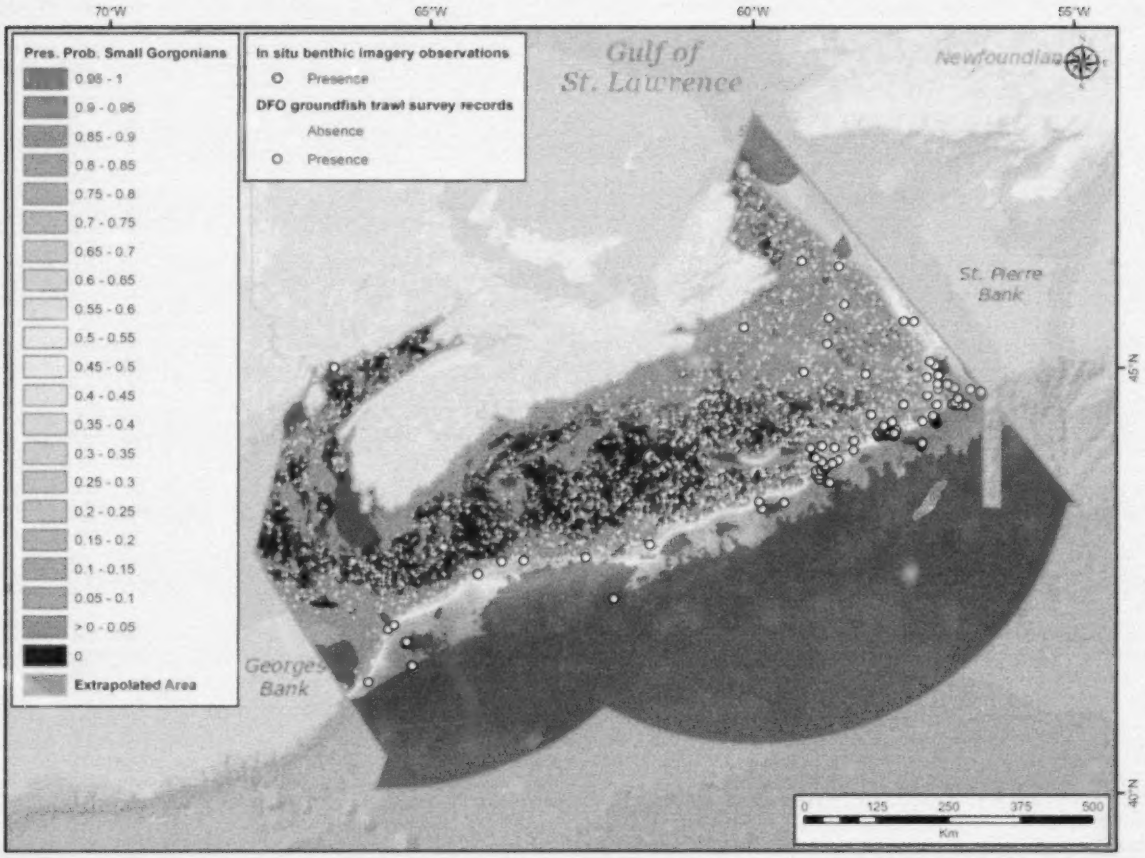

Figure 104. Areas of extrapolation of the random forest model on unbalanced presence and absence small gorgonian coral catch data from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1997 and 2014. Also shown are the presence and absence observations and predictions of presence probability (Pres. Prob.).

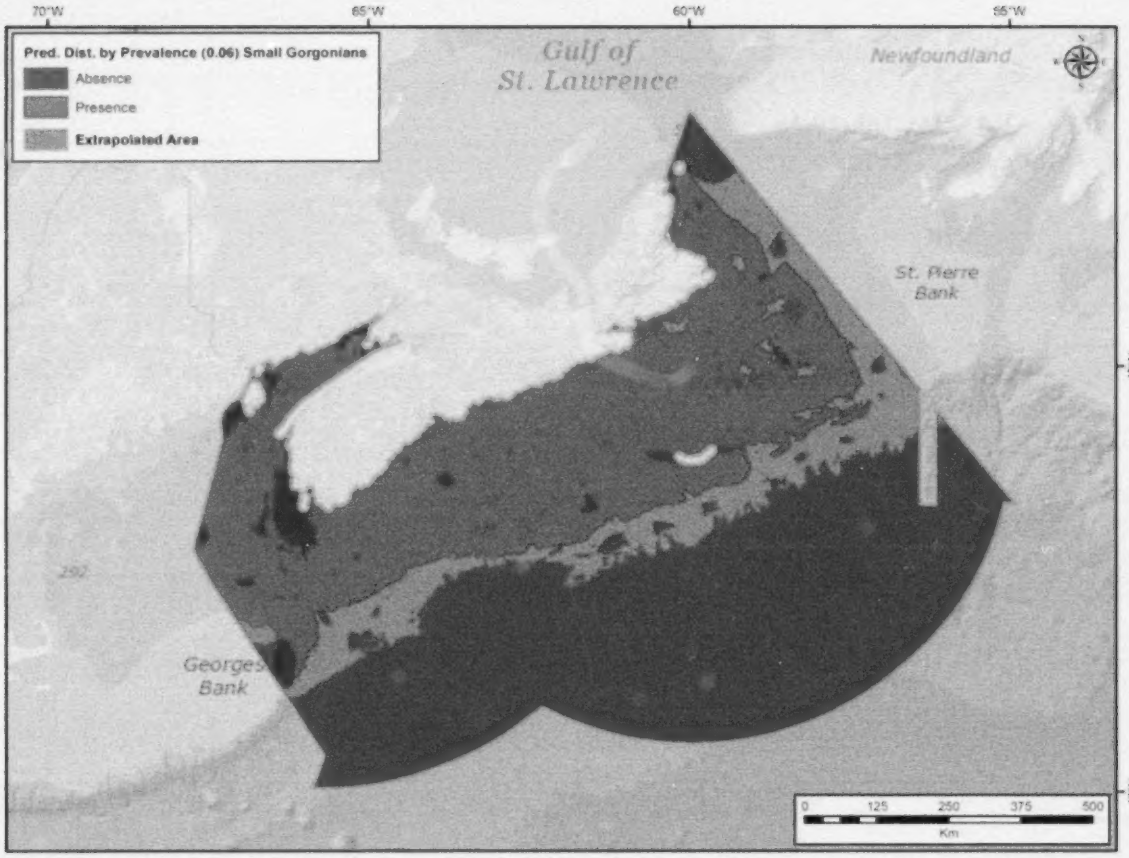

Figure 105. Predicted distribution (Pred. Dist.) of small gorgonian corals in the Maritimes Region based on the prevalence threshold of 0.06 of small gorgonian coral presence and absence data used in Model 3. Also shown are the areas of model extrapolation (grey polygon may appear red or blue).

Like in Model 2, the most important environmental predictor variable for the classification of the small gorgonian coral presence and absence data was Depth (Figure 106). This was followed more distantly by Slope, Bottom Salinity Average Range, and the remaining variables in the model. Bottom and surface salinity variables ranked high in this model. Partial dependence plots of the top 6 environmental variables are shown in Figure 107. Probability of presence of small gorgonians rapidly increased at ~300 m along the Depth gradient, and reached a plateau at ~500 m. A similar pattern was shown along the Slope gradient, where presence probability increased rapidly beginning at ~5° and plateaued at ~10 °. Presence probability was highest at the lowest Bottom Salinity Average Range and Bottom Temperature Average Range values, and highest at the highest Bottom Salinity Average Minimum and Bottom Salinity Mean values.

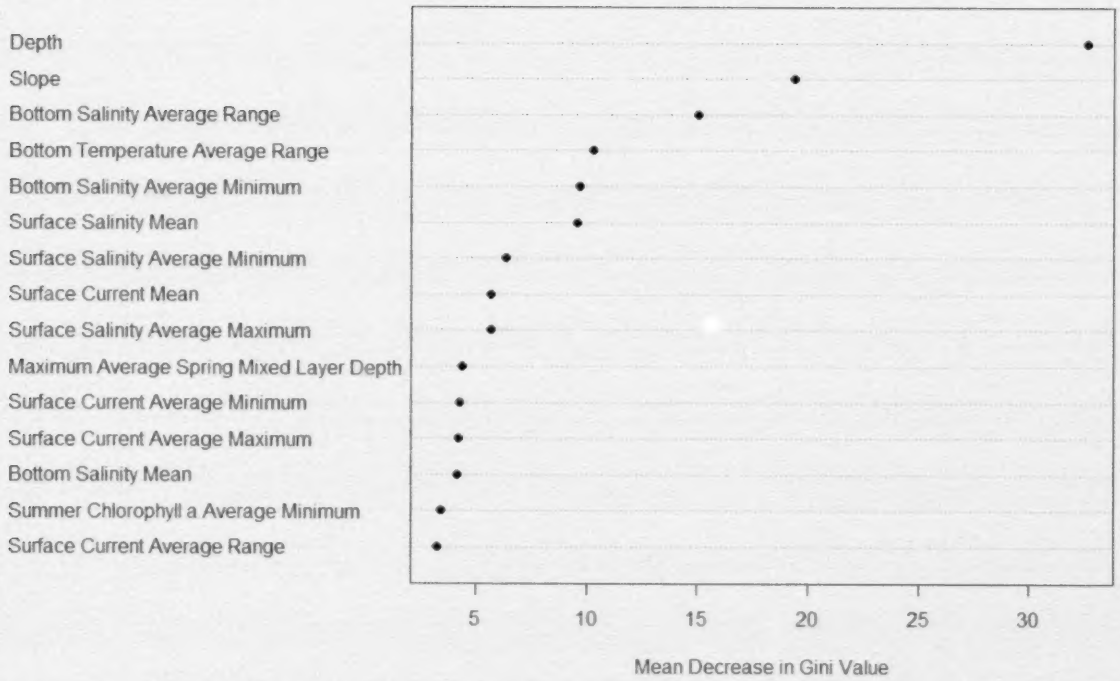

Figure 106. Importance of the top 15 predictor variables measured as the Mean Decrease in Gini value of the random forest model on unbalanced small gorgonian coral presence and absence data collected from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1997 and 2014. The higher the Mean Gini value the more important the variable is for predicting the response data.

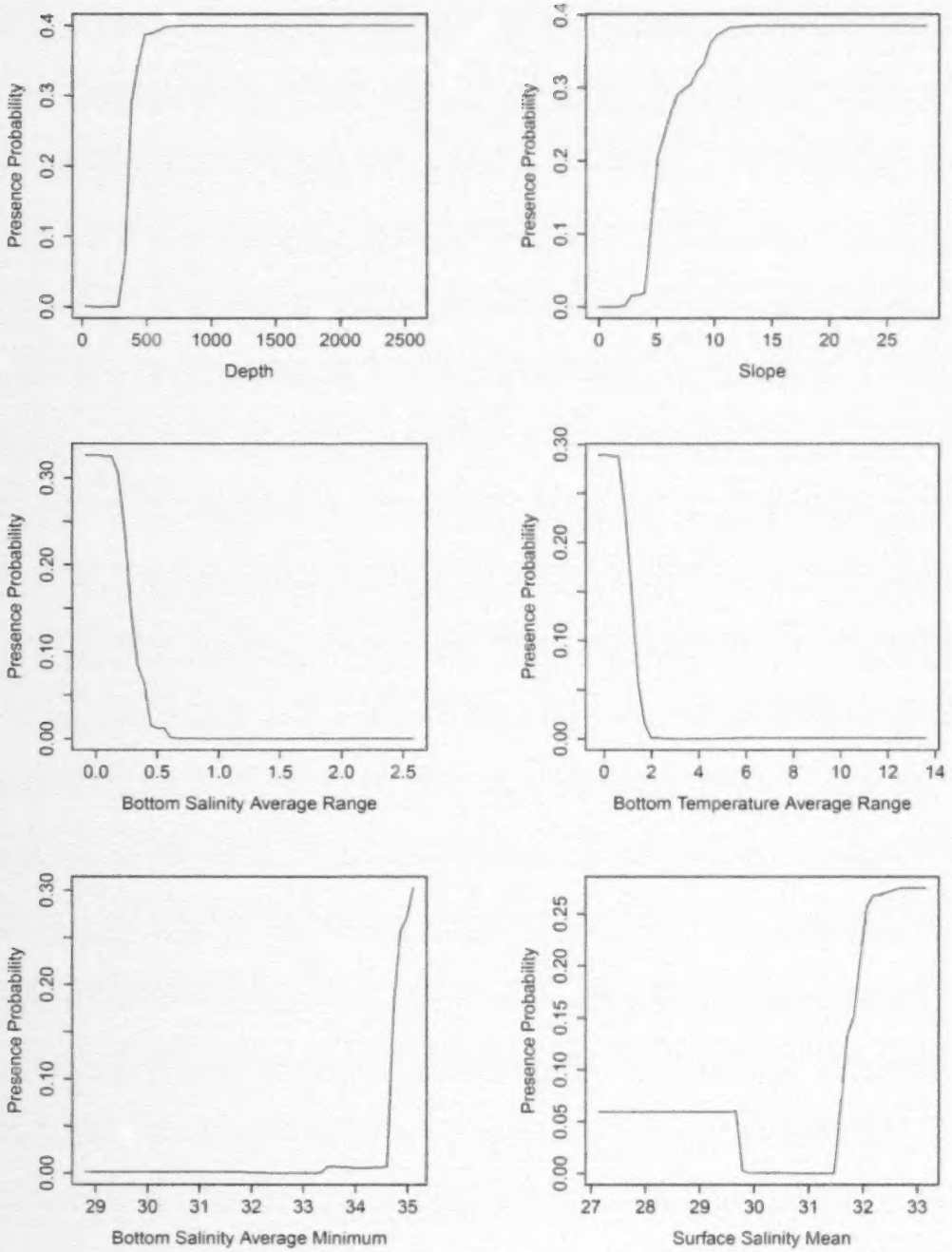

Figure 107. Partial dependence plots of the top six predictors from the random forest model of small gorgonian coral unbalanced presence and absence data collected from DFO multispecies trawl surveys, and DFO and NRCan scientific surveys conducted within the Maritimes Region between 1997 and 2014, ordered left to right from the top. Presence probability is shown on the *y*-axis.

Model Selection

The random forest model using all available small gorgonian coral records and an unbalanced species prevalence (Model 3) was selected as the best predictor of small gorgonian coral distribution in the Maritimes Region (Figure 102). Model 1 (balanced species prevalence) was considered a poor predictor of presence probability of this group due to its exaggeration of high presence probability on the banks of the eastern Scotian Shelf and in Laurentian Channel. This phenomenon was likely due to random down-sampling of the absence data. Model 2, which was generated using the same presence-absence dataset but using all absence data, produced a much more realistic presence probability surface with less exaggeration beyond the location of presence points. The additional presence records added to Model 3 produced the highest AUC and sensitivity and specificity measures of all three models. Although the presence probability surface was similar to that of Model 2, this model predicted higher small gorgonian presence probability along the eastern Scotian Shelf slope and in its canyons, and is a more accurate depiction of the spatial distribution of small gorgonians based on their known distribution in the region.

Validation of Selected Model Using Independent Data

Figure 108 shows the predicted presence probabilities of small gorgonians generated from Model 3 at the location of small gorgonian records from the NOAA Deep-Sea Coral Data Portal. Many of the NOAA records were concentrated along the Scotian Slope and in deep-water canyons, and in the eastern Gulf of Maine in shallow water. Of the 60 small gorgonian records from the NOAA data portal, 15% were predicted as absence based on the prevalence threshold of 0.06 (yellow symbols in Figure 108). The majority of these were located in shallow water in the eastern Gulf of Maine and on the Scotian Shelf. No NOAA records exist in the area of high small gorgonian presence probability in the Laurentian Channel. Several records occurred in deeper waters off the shelf in an area considered as extrapolated by the model.

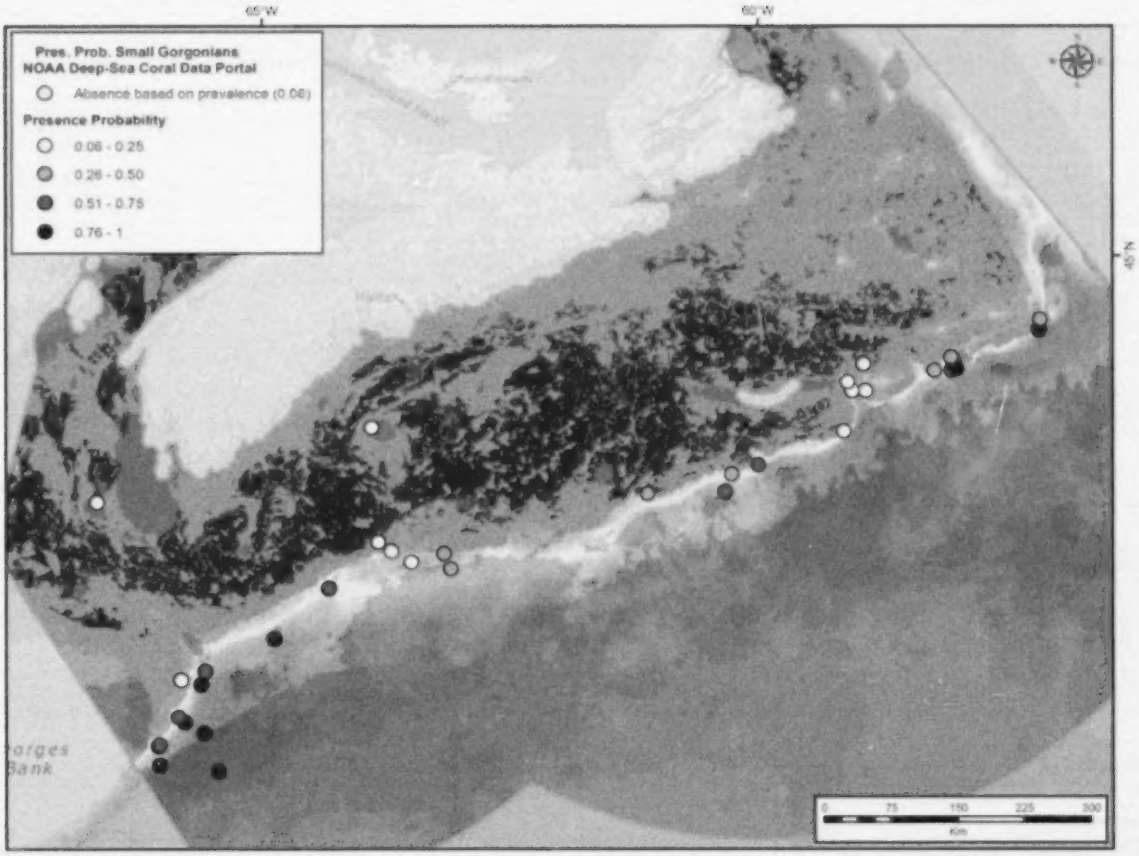

Figure 108. Validation of small gorgonian coral presence probability from Model 3 using independent data. Presence probability values were extracted to the location of small gorgonian records from the NOAA Deep-Sea Coral Data Portal.

Prediction of Small Gorgonian Biomass Using Random Forest

The accuracy measures of the regression random forest model on mean small gorgonian coral biomass per grid cell from DFO multispecies trawl surveys are presented in Table 30. The highest R^2 value was 0.423, while the average was 0.135 ± 0.155 SD. The average Normalized Root-Mean-Square Error (NRMSE) was 0.027 ± 0.0.019 SD. An R^2 value was not generated for several model folds. The percent variance explained for each fold was negative, indicating that the model had little to no predictive power.

Figures 109 and 110 show the predicted biomass surface of small gorgonians. The majority of the spatial extent was predicted to have low (0 – 0.009 kg) small gorgonian biomass. The area between Emerald and LaHave Banks had the highest predicted biomass (up to 0.15 kg). The majority of the Laurentian Channel was predicted to have low to moderate biomass of small gorgonian corals, with moderate values occurring in the southwest portion (Figure 110).

Table 30. Accuracy measures from 10-fold cross validation of a random forest model of average small gorgonian coral biomass (kg) per grid cell recorded from DFO multispecies trawl surveys in the Maritimes Region. RMSE = Root-Mean-Square Error; NRMSE = Normalized Root-Mean-Square Error.

Model Fold	R^2	RMSE	NRMSE	Percent (%) variance explained
1	3.568×10^{-5}	0.011	0.034	-11.65
2	0.212	0.007	0.022	-18.97
3	0.268	0.005	0.015	-14.21
4	N/A	0.002	0.006	-11.41
5	0.020	0.009	0.027	-14.21
6	0.030	0.025	0.073	-6.67
7	N/A	0.002	0.007	-13.21
8	0.423	0.013	0.038	-13.56
9	0.129	0.007	0.022	-15.11
10	9.692×10^{-6}	0.009	0.027	-15.86
Mean	**0.135**	**0.009**	**0.027**	**-13.49**
SD	**0.155**	**0.006**	**0.019**	**3.23**

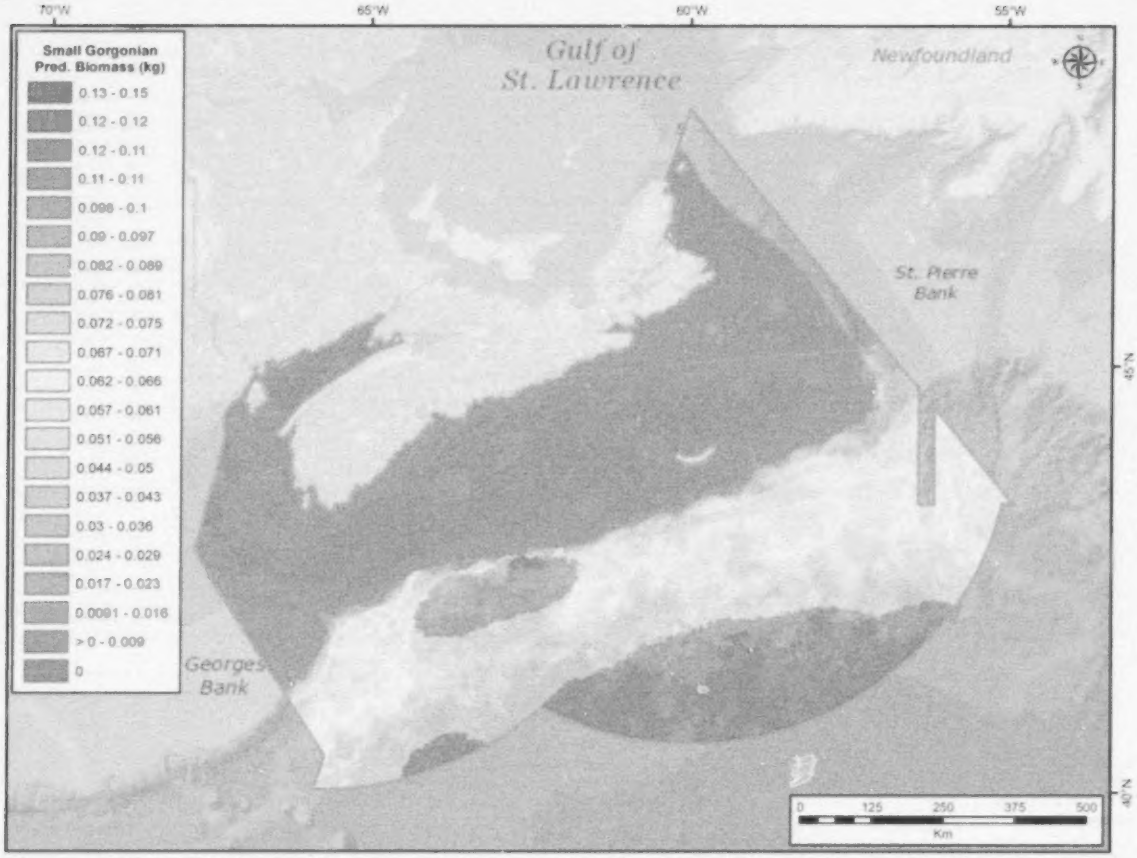

Figure 109. Predictions of biomass (kg) per grid cell of small gorgonian corals from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2002 and 2014.

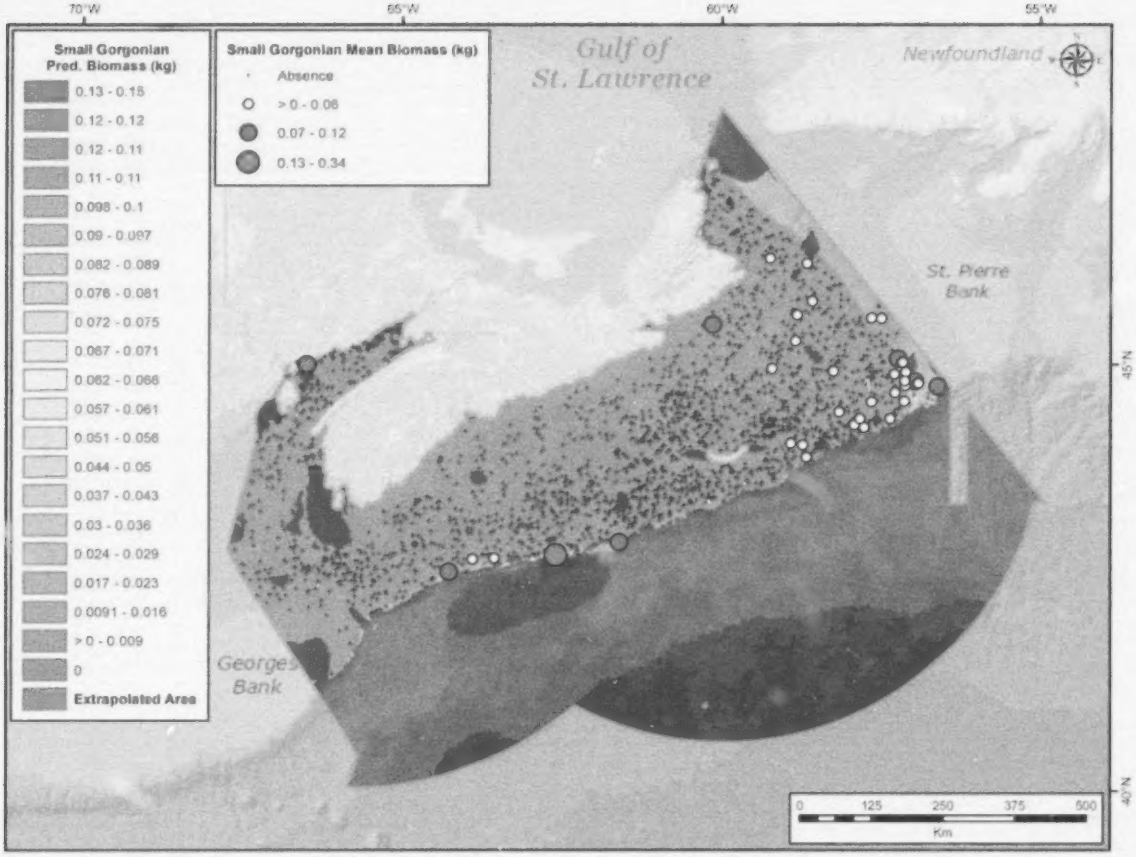

Figure 110. Predictions of biomass (kg) per grid cell of small gorgonians from catch data recorded in DFO multispecies trawl surveys conducted in the Maritimes Region between 2002 and 2014. Also shown are the mean biomass values per grid cell and areas of model extrapolation.

The top 15 most important environmental variables for predicting small gorgonian biomass are shown in Figure 111. Bottom Current Average Minimum was the most important variable in the model. Prior to spatial interpolation, this variable displayed a right-skewed distribution (Beazley et al., in prep). Examination of the Q-Q plot revealed no spatial pattern to data points over- and under-predicted by a normal distribution. Bottom Current Average Minimum was followed closely by Bottom Shear Average Minimum and Summer Chlorophyll *a* Mean. The partial dependence of small gorgonian coral biomass on the top 6 most important variables is shown in Figure 112. Predicted biomass was highest at Bottom Current Average Minimum values of 0.007 m s^{-1} and greater. Values in this range coincided with both over- and under-predicted values located across the Scotian Shelf and in the deeper portion of the study

area. The fit between predicted and observed values in the kriging model was relatively poor, with under-prediction of Bottom Current Average Minimum values of 0.007 m s^{-1} and greater. Some points could therefore be predicted lower than their true values and slightly outside the range of highest predicted biomass identified in the partial plot.

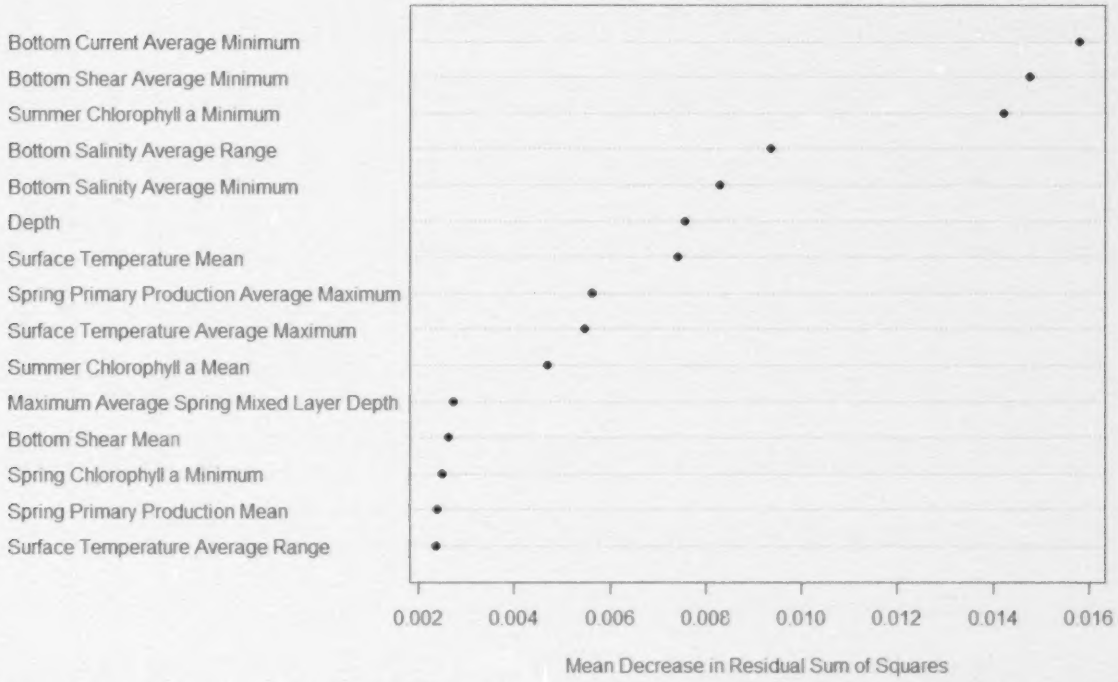

Figure 111. Importance of the top 15 predictor variables measured as the Mean Decrease in Residual Sum of Squares of the regression random forest model on small gorgonian coral mean biomass data averaged per grid cell. The higher the Mean Decrease in Residual Sum of Squares, the more important the variable is for predicting the response data.

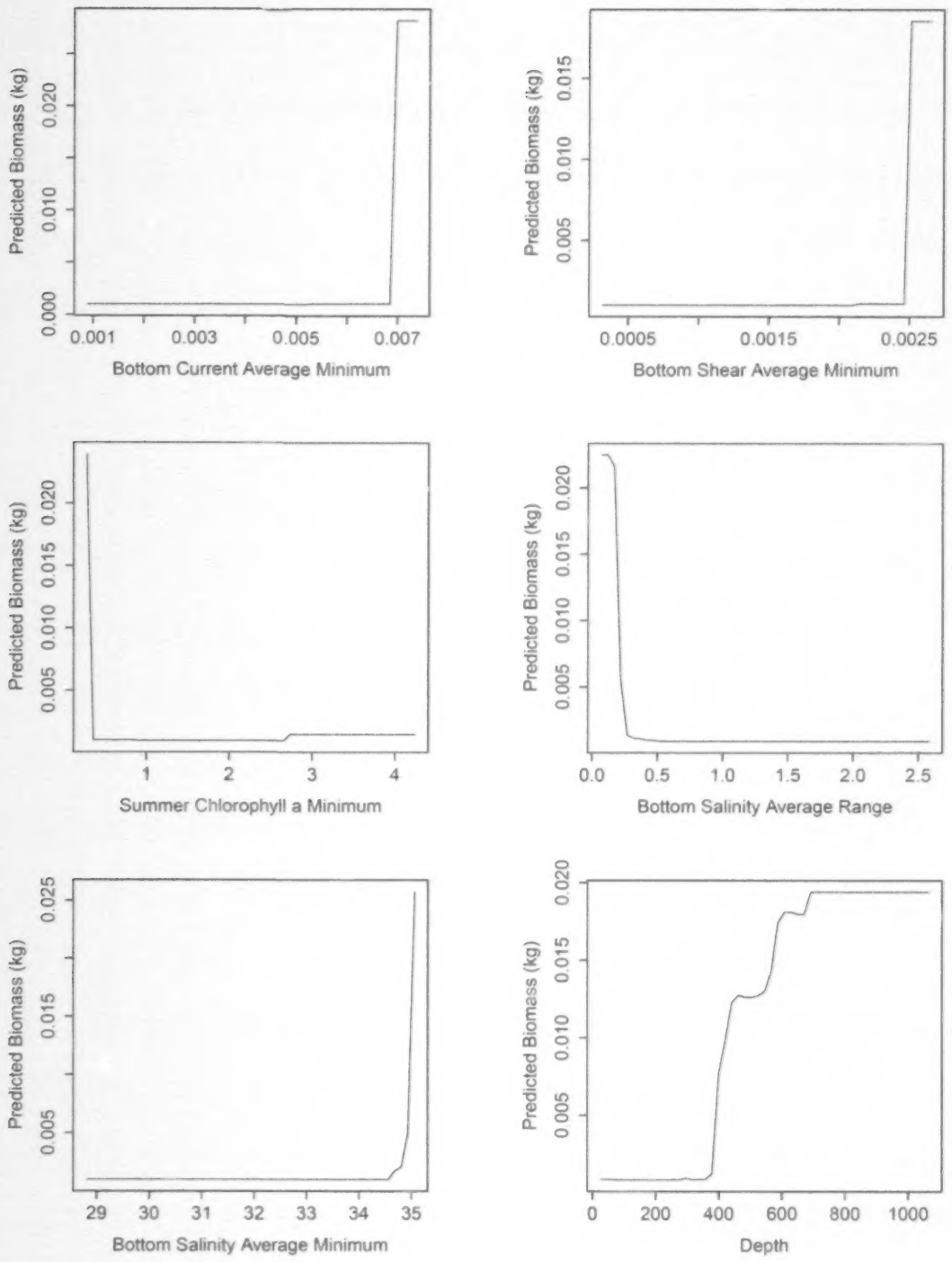

Figure 112. Partial dependence plots of the top six predictors from the random forest model of small gorgonian coral biomass data collected within the Maritimes Region, ordered left to right from the top. Predicted biomass (kg) is shown on the *y*-axis.

DISCUSSION

This study is the first to use random forest modelling to predict the distribution of several sensitive deep-sea benthic invertebrate groups in the DFO Maritimes Region. Table 31 shows a summary of the accuracy measures for the selected presence-absence and biomass random forest models of each taxonomic group. Presence-absence models for most groups performed very well, with cross-validated AUC values ranging from 0.760 to 0.977. The poorer performance of the sponge model compared to the others is likely due to the coarse taxonomic resolution of this group (phylum level), and the inclusion of a number of different sponge species with a preference for different environmental conditions (e.g. shelf and slope species). Species with restricted niches and higher habitat specificity are commonly modelled with higher accuracy than generalist species (McPherson and Jetz, 2007; Tsoar et al. 2007). This was also found in our study, as the highest performing model was of a single species, *Vazella pourtalesi*, which had a narrow spatial distribution and a high affinity for areas with high (>34) bottom salinity. Previous RF models were performed on Porifera using a biomass threshold which served to distinguish the *Geodia*-dominated habitats on the continental slopes from the smaller shelf species (Knudby et al., 2013). Those models performed well but were aided by the presence of the large massive sponges and their associated sponge fauna which allowed a habitat (rather than just Porifera) to be modelled. On the Scotian Shelf, *V. pourtalesi* is one of the larger sponges present and it is not strongly associated with other sponge species. The relatively good performance of the sea pens, and large and small gorgonian coral models, despite the inclusion of multiple species in these groups, could be attributed to a narrow range tolerance and preference for similar environmental conditions between species (Bryan and Metaxas, 2007), and to their lower position in the taxonomic hierarchy with fewer species included in each taxon. Good predictive performance of species distribution models on coral species aggregated at higher taxonomic levels was also noted by Lagasse et al. (2015). In general, corals are predicted to occur in areas with complex topography and strong currents (Bryan and Metaxas, 2007). However, Bryan and Metaxas (2007) noted that despite a similar overall preference for certain environmental conditions, different families of coral (Paragorgiidae and Primnoidae) preferred slightly different combinations of environmental variables from one another, suggesting the importance of modelling individual species as opposed to aggregated taxonomic groups. In this study, modelling some species individually was not feasible due to the low number of occurrences from validated sources over the spatial extent. Nonetheless, these results highlight the need for increased data collection and improved taxonomic identification, particularly of sponge invertebrate catch onboard both commercial and research vessels.

Unlike the classification models on presence-absence data, performance of the regression random forest models on biomass did not appear to follow any trends relating to the taxonomic resolution of the response data, with poor performance of all models except those on sea pens

Table 31. Summary of the mean accuracy measures for selected presence-absence models and biomass models for each of the five taxonomic groups. NRMSE = Normalized Root-Mean-Square-Error.

	Presence-absence			Biomass	
	AUC	**Sensitivity**	**Specificity**	**R^2**	**NRMSE**
Sponges (Porifera)	0.760	0.691	0.702	0.130	0.030
Vazella pourtalesi	0.977	0.952	0.913	0.087	0.024
Sea Pens (Pennatulacea)	0.901	0.813	0.819	0.518	0.018
Large Gorgonian Corals	0.928	0.833	0.892	0.285	0.016
Small Gorgonian Corals	0.949	0.876	0.916	0.135	0.027

and large gorgonian corals (see Table 31). There was little extrapolation of high biomass predictions beyond the location of the highest catches in the raw data, and predicted biomass was overall much lower than the empirical maximum for each taxonomic group. The random forest and generalized additive models (GAMs) predicted similar areas of high biomass of the coral and sponge groups (see Appendix 1). For some groups (e.g. small gorgonian corals, see Figures A1.8 and A1.9), GAMs provided better predictions of biomass along the slope and in Laurentian Channel compared to the random forest model. The poor performance of the models for sponges and *V. pourtalesi* may reflect an effect of previous fishing. Biomass data are more sensitive to the effects of fishing than presence-absence data both in terms of catchability and in the inaccurate reflection of virgin biomass presented on fished shelves. Comparison of the fishing history of the area with the distributions of each taxon may help to explain these results. In a recent application of random forest to VME indicator taxa biomass in the NAFO Regulatory Area (see Downie et al., 2016), fishing intensity was used as a covariate, which helped to explain some of the observed patterns. This is something that could be done in the future with our models to explore the effect of fishing history on biomass. Sea pens are less vulnerable to fishing disturbance than some other taxa modelled in this report, and the areas with large gorgonian corals are often avoided as they prefer rocky substrates (Kenchington et al., 2011). Alternatively, poor model performance for the sponges and the small gorgonian corals also may have resulted from a highly imbalanced design and large number of zero catches (absences) included in the models, and/or high variability in the positive catches (Li and Heap, 2008). This is particularly likely for the small gorgonian corals. High variability and low data density of the response has shown to negatively affect regression random forest model performance estimates (Bučas et al., 2013). However, Newbold et al. (2009) suggest that robust sampling across the full range of environmental gradients over sampling density alone is the most important determinant of model accuracy, highlighting the importance of equal and unbiased samples for species distribution modelling.

The models based on presence-absence data worked well at interpolating predictions between data observations and extrapolating within the shallow (< 2000 m depth) portion of the

study extent. However, the Maritimes Region extends out to the Canadian EEZ to approximately 5100 m water depth. We are not confident of the model extrapolations to those depths as we have no means of validation. Sponges, sea pens and gorgonian corals can be found at such depths and so the model may be helpful in guiding research surveys to perform such validation. For biomass, the GAMs did not serve to resolve predictions in these extrapolated areas. An exception was the sponge GAM model, which predicted a localized area of high biomass on the Scotian Rise.

We have found that classification random forest models generated using all presence and absence data (i.e. unbalanced species prevalence) and a threshold equal to species prevalence produced the most realistic presence probability prediction surfaces and highest model accuracy in instances when the input data were highly imbalanced and spatially biased across the study area. Random down-sampling of the absence data often resulted in gross extrapolation of high presence probability beyond the location of presence observations. This was likely exacerbated when down-sampling to match a low number of presence observations, as in our *V. pourtalesi*, sea pen, and gorgonian coral models. Our sponge model however, produced nearly identical presence probability surfaces and model accuracy measures between balanced and unbalanced runs, likely due to the high and relatively even number of presence and absence observations across the study extent. These results may help guide future applications of random forest modelling by providing insight into which methods are appropriate based on the properties of the training data.

The use of records from different data sources (e.g. multispecies trawl surveys and *in situ* camera observations) in random forest modelling may introduce bias and cause poor model performance, particularly if there are notable differences in catchability (i.e., the ability to detect a presence) between gear types. Catchability for corals and sponges is unknown for research vessel trawl gear but is assumed to be low for some species. In the NAFO Regulatory Area gear efficiency of Campelen and Lofoten trawl gear for large-sized sponges was estimated at 2%, although others reported up to 70% (see review in Kenchington et al., 2011). In contrast, there is high detectability from *in situ* photos and video footage, although a smaller width of seabed is surveyed compared to trawl gear. Many of DFO's scientific missions involving benthic imagery collection were designed to target the continental slope and canyons where deep-water corals are known to congregate. These areas are typically not surveyed in the multispecies stock assessment surveys as they are either outside of the survey depth limit or are too rough to deploy bottom-tending gear. The addition of *in situ* camera observations and other sources significantly improved the predictive performance of the presence-absence models, and its inclusion in the models is warranted given the spatial bias in the DFO multispecies trawl surveys. Naturally, the addition of the *in situ* camera observations increased the probability of occurrence of the three coral groups along the Scotian Slope and in several deep canyons. This increase was most noticeable for sea pens and small gorgonians, where high predictions of presence probability were nearly extended across the width of the continental slope. For large gorgonians there was

little extrapolation of high presence probability beyond the deepest canyons (the Northeast Channel and The Gully) where the camera observations were concentrated. Nonetheless, predictions in areas dominated by *in situ* camera observations should be interpreted with caution, as no absence data accompanied those records.

Although our goal was not to identify the specific niche requirements of each coral and sponge group, ecological interpretation of the random forest models is necessary for determining the validity of the predictions and whether they are consistent with known information on distribution and biology. Corals were predicted to have the highest presence probability along the slope and in deep-water canyons, results that are consistent with the known distribution of these organisms (Breeze et al., 1997; Gordon and Kenchington, 2007; Cogswell et al., 2009). Depth and Slope were the top two predictors of the sea pen, and large and small gorgonian presence-absence data. Sea pens and small gorgonians had similar presence probability surfaces and were predicted to occur over a broader range than large gorgonian corals. All sea pens and the majority of small gorgonians anchor in soft substrate, whereas large gorgonians are restricted to areas containing cobble, boulder, or large rocky outcrops. Substrate, which is considered a major limiting factor for the distribution of deep-water corals (Bryan and Metaxas, 2007) was not included in the models. There is currently no high-resolution substrate layer available for the Maritimes Region. Point-source sediment grain size data are readily available for the region through Natural Resources of Canada's (NRCan) public domain (http://ed.gdr.nrcan.gc.ca/index_e.php), however, spatial interpolation of this data for the creation of a continuous layer requires a high density of observations in order to accurately capture differences in sediment characteristics that occur on the micro-scale. Further, habitat heterogeneity over much of the modelling domain would occur over scales smaller than the resolution of our environmental data (< 1 km; cf. Cuff et al., 2015; Rincón and Kenchington, 2016). In the absence of substrate, other variables included in our models such as slope and bottom shear may be considered as proxies for substrate. Partial dependence plots showed that large and small gorgonians were restricted to areas with a greater slope than sea pens (10° and greater for large and small gorgonians versus 5° and greater for sea pens). Nonetheless, the full environmental range of these species has not been sampled, as indicated by partial dependence plots that do not reach zero presence probability at both extremes along each environmental gradient (Knudby et al., 2013). Future sampling conducted over a greater geographic and depth range would provide more information from which more accurate physiological tolerance limits could be ascertained.

In summary, the excellent predictive capacity of most presence-absence models show that the spatial distribution of these taxonomic groups supports the use of random forest species distribution models for the prediction of significant benthic invertebrate habitat in the Maritimes Region. However, we have found that properties of the response data can greatly affect model performance. In our study, a single-species model performed better than those on aggregated species, particularly the sponges that were aggregated at the phylum level.

In the absence of data observations, the results of this study could be used to identify the potential distribution of benthic EBSA/VME indicator taxa for use in fisheries management and conservation applications. For instance, we have shown that the potential habitat of the Russian Hat sponge *Vazella pourtalesi* far exceeds the boundaries of two areas closed for the protection of this species from bottom-tending gear in Emerald Basin.

The SDMs generated in this report identify potential species distribution and can indicate areas for future restoration initiatives towards the implementation of the Policy for Managing the Impact of Fishing on Sensitive Benthic Areas. This policy was developed by DFO in 2009 to ensure Canadian fisheries are conducted in a manner that supports marine conservation and sustainable resource use within and outside Canada's 200 nautical mile EEZ. These models provide continuous surfaces of presence and biomass that can fill in gaps in survey coverage and extrapolate to a certain degree to areas outside of the surveys. Combined with kernel density analysis (Kenchington et al., 2016), SDM can be used to refine significant benthic area polygons produced by the former by clipping boundaries to more probabilistic borders.

ACKNOWLEDGMENTS

We thank Marty King and Derek Fenton (DFO) for their guidance throughout the development of this report, and in particular for their suggestions on useful maps for ocean managers. We also thank Marty King and Ben Lowen (DFO) for their review of this report. This project was funded in part by a one year project under DFO's Strategic Program for Ecosystem-Based Research and Advice (SPERA) to EK and through financial support by DFO's Oceans and Coastal Management Division, Ecosystem Management Branch at the Bedford Institute of Oceanography to AK.

REFERENCES

Beazley, L., Lirette, C., Sabaniel, J., Wang, Z., Knudby, A., and Kenchington, E. 2016. Characteristics of Environmental Data Layers for Use in Species Distribution Modelling in the Gulf of St. Lawrence. Can. Tech. Rep. Fish. Aquat. Sci. 3154: viii + 357p.

Breeze, H., Davis, D.S., Butler, M., and Kostylev, V. 1997. Distribution and Status of Deep Sea Corals off Nova Scotia. Ecology action Centre, Marine Issues Committee Species Publication 1: 1-58.

Breiman, L. 2001. Random forests. Machine Learning, 45: 5–32.

Breiman, L., Friedman, J.H., Olshen, R. and Stone, C.J. 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific California.

Bryan, T.Y. and Metaxas, A. 2006. Predicting suitable habitat for deep-water gorgonian corals on the Atlantic and Pacific Margins of North America. Mar. Ecol. Prog. Ser. 330: 113-126.

Bučas, M., Bergström, U., Downie, A-L., Sundblad, G., Gullström, M., von Numers, M., Šiaulys, A., and Lindegarth, M. 2013. Empirical modelling of benthic species distribution, abundance, and diversity in the Baltic Sea: evaluating the scope for predictive mapping using different modelling approaches. ICES J. Mar. Sci. doi: 10.1093/icesjms/fst036.

CBD. 2009. Ninth meeting of the Conference of the Parties to the Convention on Biological Diversity, 19-30 May 2008 - Bonn, Germany. Marine and Coastal Biodiversity, (COP) 9, Decision IX/20.

Chen, X., and Ishwaran, H. 2012. Random forests for genomic data analysis. Genomics 99: 323–329.

Chen, C., Liaw, A., and Breiman, L. 2004. Using random forest to learn imbalanced data. Berkeley: University of California.

Cogswell, A.T., Kenchington, E.L.R., Lirette, C.G., MacIsaac, K., Best, M.M., Beazley, L.I., and Vickers, J. 2009. The current state of knowledge concerning the distribution of coral in the Maritimes Provinces. Can. Tech. Rep. Fish. Aquat. Sci. 2855, v + 66pp.

Cuff, A., Anderson, J.T., and Devillers, R. Comparing surficial sediments maps interpreted by experts with dual-frequency acoustic backscatter on the Scotian shelf, Canada. Cont. Shelf Res. 110: 149–161.

DFO, 2004. Identification of Ecologically and Biologically Significant Areas. DFO Can. Sci. Advis. Sec. Ecosystem Status Rep. 2004/006.

DFO. 2014. Offshore Ecologically and Biologically Significant Areas in the Scotian Shelf Bioregion. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2014/041.

Downie, A.-L., Kenny, A., and Barrio-Frojan, C. 2016. Predictive models of VME indicator taxa biomass in the NAFO Regulatory Area- including the effects of fishing activity. NAFO SCR Doc., in prep.

Drinkwater, K.F., Petrie, B., Smith, P.C. 2002. Hydrographic variability on the Scotian Shelf during the 1990s. NAFO SCR Doc. 02/42, Serial No. N4653.

Dunn, P. K., and Smyth, G. K. 1996. Randomized quantile residuals. Journal of Computational and Graphical Statistics, 5: 236-244.

Elith, J., Kearney, M., and Phillips, S. 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1: 330-342.

ESRI. 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Evans J.S., Murphy, M.A., Holden, Z.A., and Cushman, S.A. 2011. Modeling Species Distribution and change Using Random Forests. In: Predictive Species and Habitat Modeling in Landscape Ecology: Concepts and Applications. Eds: Drew, C.A., Wiersma, Y.F., and Huettmann, F. Springer, NY.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recog. Lett. 27: 861-874.

Gass, S.E. 2002. An Assessment of the Distribution and Status of Deep Sea Corals in Atlantic Canada by Using both Scientific and Local Forms of Knowledge. MES thesis, Dalhousie University, Halifax, Canada.

Gordon, D.C. and Kenchington, E.L.R. 2007. Deep-Water Corals in Atlantic Canada: A Review of DFO Research (2001-2003). Proc. N.S. Inst. Sci. 44: 27-50.

Hastie, T., and Tibshirani, R. 1986. Generalized additive models. Stat. Sci. 1: 297-318.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. 2005. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. Springer+Verlag.

Hanberry, B.B. and He, H.S. 2013. Prevalence, statistical thresholds, and accuracy assessment for species distribution models. Web Ecol. 13: 13-19.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. 2005. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Second Edition. Springer+Verlag.

Herrick, K.K., Huettmann, F., and Lindgren, M.A. 2013. A global model of avian influenza prediction in wild birds: the importance of northern region. Vet. Res. 44:42.

Jiménez-Valverde, A. and Lobo, J.M. 2006. The ghost of unbalanced species distribution data in geographical model predictions. Diversity Dist. 12: 521-524.

Kenchington, E. 2014. A General Overview of Benthic Ecological or Biological Significant Areas (EBSAs) in Maritimes Region. Can. Tech. Rep. Fish. Aquat. Sci. 3072: iv + 45p.

Kenchington, E., Lirette, C., Cogswell, A., Archambault, D., Archambault, P., Benoit, H., Bernier, D., Brodie, B., Fuller, S., Gilkinson, K., Lévesque, M., Power, D., Siferd, T., Treble, M., and Wareham, V. 2010. Delineating Coral and Sponge Concentrations in the Biogeographic Regions of the East Coast of Canada Using Spatial Analyses. DFO Can. Sci. Advis. Sec. Res. Doc. 2010/041. vi + 202pp.

Kenchington, E., Murillo, F.J., Cogswell, A., and Lirette, C. 2011. Development of Encounter Protocols and Assessment of Significant Adverse Impact by Bottom Trawling for Sponge Grounds and Sea Pen Fields in the NAFO Regulatory Area. NAFO SCR Doc. 11/75, Serial No. N6005.

Kenchington, E., Lirette, L., Murillo, F.J., Beazley, L., Guijarro, J., Wareham, V., Gilkinson, K., Koen Alonso, M., Benoit, H., Bourdages, H., Sainte-Marie, B., Treble, M., Siferd, T. 2016. Kernel Density Analyses of Coral and Sponge Catches from Research Vessel Survey Data for Use in Identification of Significant Benthic Areas. Can. Tech. Rep. Fish. Aquat. Sci. 3167: viii + 206p.

Knudby, A., Kenchington, E., Murillo, F.J. 2013. Modelling the distribution of *Geodia* sponges and sponge grounds in the Northwest Atlantic. PLoS One 8, e82306. http://dx.doi.org/10.1371/journal.pone.0082306.

Kuhn, M. and Johnson, K. 2013. Applied Predictive Modeling. New York: Springer Science + Business Media.

Lagasse, C.R., Knudby, A., Curtis, J., Finney, J.L., and Cox, S.P. 2015. Spatial analyses reveal conservation benefits for cold-water corals and sponges from small changes in a trawl fishery footprint. Mar. Ecol. Prog. Ser. 528: 161-172.

Li, J. and Heap, A.D. 2008. A Review of Spatial Interpolation Methods for Environmental Scientists. Geoscience Australia, Record 2008/23, 137 pp.

Liaw, A. and Wiener, M. 2002. Classification and regression by randomForest. R News 2, 18-22.

Liu, C., Berry, P.M., Dawson, T.P., and Pearson, R.G. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28: 385-393.

Matthiopoulos, J. 2011. How to be a Quantitative Ecologist: The 'A to R' of Green Mathematics and Statistics. Wiley, Chichester, West Sussex.

McPherson, J.M., Jetz, W., and Rogers, D.J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41: 811-823.

McPherson, J.M. and Jetz, W. 2007. Effects of species' ecology on the accuracy of species distribution models. Ecography 30: 135-151.

Miller, D.L., Rexstad, E., Burt, L., Bravington, M.V., and Hedley, S. 2015. Package 'dsm'. 26 p.

Mortensen, P.B., Buhl-Mortensen, L., Gass, S.E., Gordon Jr., D.C., Kenchington, E.L.R., Bourbonnais, C., and MacIsaac, K. 2006. Deep-water Corals in Atlantic Canada: A Summary of ESRF-Funded Research (2001-2003). Environmental Studies Research Funds (Canada) 143: 1-83.

NAFO. 2013. Conservation and Enforcement Measures. NAFO/FC, Doc. 13/1, Serial No. N6131. 103p.

Newbold, T., Reader, T., Zalat, S., El-Gabbas, A., and Gilbert, F. 2009. Effect of characteristics of butterfly species on the accuracy of distribution models in an arid environment. Biodivers. Conserv. 18: 3629-3641.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rincón, B and Kenchington, E. 2016. Spatial and temporal variation of benthic macrofauna on the eastern Scotian Shelf: association with juvenile Haddock (*Melanogrammus aeglefinus*) spatial structure and environmental drivers. Under Review.

Shono, H. 2008. Application of the Tweedie distribution of zero-catch data in CPUE analysis. Fish. Res. 93: 154–162.

Tremblay, M.J., Black, G.A.P., and Branton, R.M. 2007. The distribution of common decapod crustaceans and other invertebrates recorded in annual ecosystem surveys of the Scotian Shelf 1999-2006. Can. Tech. Rep. Fish. Aquat. Sci. 2762, iii + 74p.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., and Kadmon, R. 2007. A comparative evaluation of presence-only methods for modelling species distribution. Biod. Res. 13: 397-405.

Tweedie, M.C.K. 1984. An index which distinguishes between some important exponential families. In: Ghosh, J.K., Roy, J. (Eds.), Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Indian Statistical Institute, Calcutta, pp. 579–604.

Wood, S.N. 2006. Generalized additive models: an introduction with R. Chapman & Hall/CRC Press, Boca Raton, FL.

APPENDIX 1

Alternative Prediction Models- Generalized Additive Models for Predicting Coral and Sponge Biomass in the Maritimes Region

Given the fair to poor prediction of biomass by the random forest models, particularly in deep water, generalized additive models (GAMs; Hastie and Tibshirani, 1986) were developed to compare to the random forest results and to determine whether predictions could be improved for the areas considered as extrapolated by random forest models. A generalized additive model (Hastie and Tibshirani, 1986; 1990) is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates. GAM models follow this general structure:

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots$$

where $\mu_i \equiv E(Y_i)$ and $Y_i \sim$ some exponential family distribution. Y_i is a response variable, X_i^* is a row of the model matrix for any strictly parametric model components, θ is the corresponding parameter vector, and the f_j are smooth functions of the covariates, x_k (Wood, 2006). The model allows for somewhat flexible specification of the dependence of the response on the covariates. This flexibility provides potential for a better fit to the data than purely parametric models.

Two different approaches were used to select the predictor variables. In the first approach, highly correlated variables were identified and eliminated in order to increase interpretability of the models and to reduce the effects of collinear variables. This was done following the variable elimination procedure outlined in Knudby et al. (2013a). The Spearman's rank correlation coefficients between all predictor variables in the study area were calculated from all raster cells in the study area, and the two predictors with the highest correlation were then considered and one of them eliminated. This process was repeated until there were no variables remaining that were correlated higher than 0.7. Models generated using the variables selected with this approach are termed 'GAM 0.7 Variables' herein. The second approach involved selecting the top predictor variables identified in the random forest biomass models. This was done independently for each taxonomic group. Those variables with a higher influence in the RF models were identified by examining the importance plots and identifying those variables that fell above a natural break in the Mean Decrease in Residual Sum of Squares. Models generated with variables selected using this approach are termed 'GAM RF Variables' herein.

The Tweedie distribution (Tweedie, 1984) was utilized for each model. The Tweedie model is an expansion of a compound Poisson model derived from the stochastic process where the weight of the counted objects has a gamma distribution. This model has the advantage of handling the zero-catch data in a unified way and has shown to outperform the two-stage Delta

lognormal model (Shono, 2008). The 'mgcv' package in R (Wood, 2006) was used to construct the GAMs.

Shrinkage smoothers were applied to each covariate in the form of a penalized cubic regression spline (s(variable, bs='cs')). Shrinkage smoothers allow the 'wiggliness' of each covariate to go to zero as required by the data (Matthiopoulos, 2011). Shrinkage smoothers are useful for variable selection, as such covariates remain in the model but have no effect on model predictions. For each model, autocorrelation in the residuals was determined by examining ACF plots. When autocorrelation appeared substantial, latitude and longitude were included in the model as a tensor product (i.e. te(lat, long)).

Model performance was evaluated by assessing the goodness-of-fit statistic R^2, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), and the percent (%) deviance explained. Accuracy measures and model performance were compared between models generated using each set of predictor variables.

Residual plots to evaluate the fitness of the model were generated using the 'gam.check' function of the mgcv package. However, an artifact of the link function shows exact zeros as a band along the residuals vs. linear predictor plot, making it difficult to see whether residuals show heteroskedasticity. In order to avoid this issue, randomized quantile residuals (Dunn and Smyth, 1996) were generated using the 'rqgam.check' function of the 'dsm' package (Miller et al., 2015). Randomized quantile residuals transform the residuals to be exactly normally distributed, therefore removing artifacts generated by the link function and making the residuals vs. linear predictor plot easier to interpret.

When predicting to the entire study extent of the Maritimes Region, the GAM models often produced erroneously-high biomass values. In these cases, additional models were tried that included latitude and longitude, and modifications to the k value (k=4). These models were evaluated against one another using the above criteria.

Sponges (Porifera)

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sponge biomass are presented in Table A1.1. Both models performed poorly, with R^2 values less than 0.04. Accuracy measures were comparable between both models, but the AIC and BIC values were slightly better for the GAM 0.7 Variables model. The significant of each predictor variable for both the GAM RF Variables and GAM 0.7 Variables models are shown in Tables A1.2 and A1.3, respectively.

Figure A1.1 shows the graphical diagnostics for both models. Both showed fairly normal residuals. The residuals vs. linear predictor plots showed patterns indicative of heteroskedasticity, while the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

The GAM 0.7 Variables model predicted erroneously-high biomass values when applied to the Maritimes Region study extent, therefore the predicted surface is not presented here. Additional models were generated with latitude and longitude and with k=4 for each predictor variable. Although performance measures were slightly higher for these models, erroneous predicted values still occurred and therefore these model results are not presented here.

Figure A1.2 shows the predicted biomass surface of sponges from the GAM RF Variables model. The majority of the study area was predicted to have zero biomass of sponge. Areas of high biomass (up to 171.57 kg) occurred along a band in deeper waters off the central Scotian Slope. This area of high biomass was not supported by the presence of data observations (Figure A1.2). The model predicted low biomass to occur in Emerald Basin where the highest sponge catches were recorded, corroborating the poor model performance indicated in the performance measures (Table A1.1) and diagnostic plots (Figure A1.1).

Table A1.1. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sponges in the Maritimes Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.043	0.035
Deviance explained	20.30%	25.20%
AIC	5689.606	5574.534
BIC	6005.015	5983.860

Table A1.2. Results of the GAM RF Variables model built to predict the biomass of sponges in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Temperature Mean	7.204	4.513	1.610×10^{-5}*
Bottom Temperature Average Minimum	4.545	5.329	3.140×10^{-5}*
Fall Chlorophyll *a* Minimum	4.530	4.692	2.980×10^{-4}*
Fall Primary Production Average Range	3.816	13.329	5.140×10^{-12}*
Summer Primary Production Average Maximum	4.318	3.819	0.001*
Summer Primary Production Mean	7.815	8.525	5.920×10^{-12}*
Summer Primary Production Average Range	6.620	12.424	7.210×10^{-16}*
Surface Current Average Range	3.534	9.589	6.220×10^{-8}*

Table A1.3. Results of the GAM 0.7 Variables model built to predict the biomass of sponges in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Minimum	1.713	16.429	5.510×10^{-8}*
Bottom Salinity Average Minimum	6.276	14.154	$< 2 \times 10^{-16}$*
Bottom Temperature Average Minimum	4.224	9.412	4.500×10^{-9}*
Annual Chlorophyll *a* Range	2.523	2.262	0.103
Spring Chlorophyll *a* Maximum	4.654	8.758	1.260×10^{-18}*
Depth	3.916	31.827	$< 2 \times 10^{-16}$*
Annual Primary Production Mean	5.976	7.996	1.210×10^{-9}*
Annual Primary Production Average Minimum	3.468	4.934	4.770×10^{-4}*
Fall Primary Production Average Maximum	6.147	8.827	6.000×10^{-11}*
Fall Primary Production Average Minimum	1.520	4.826	0.008*
Spring Primary Production Average Maximum	0.001	0.007	0.996
Spring Primary Production Average Minimum	5.821	17.832	$< 2 \times 10^{-16}$*
Surface Salinity Average Range	6.195	4.576	4.840×10^{-5}*
Slope	2.393	14.770	5.360×10^{-9}*

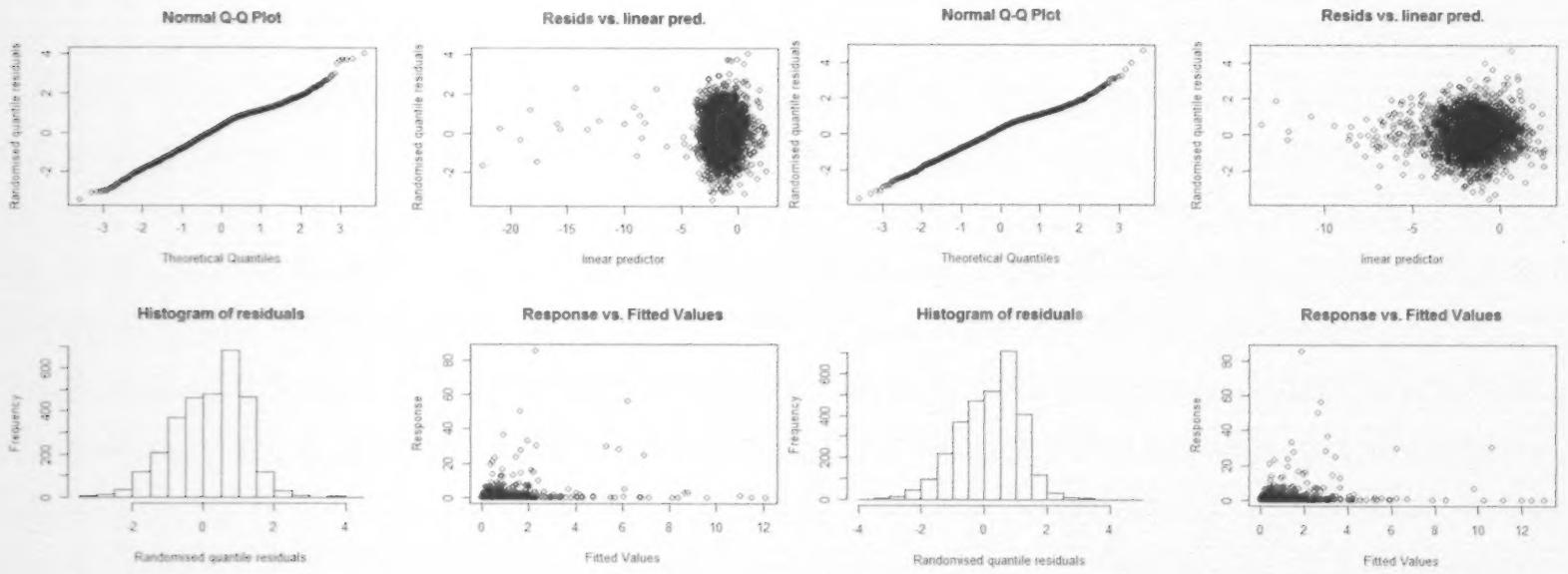

Figure A1.1. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the biomass of sponges in the Maritimes Region.

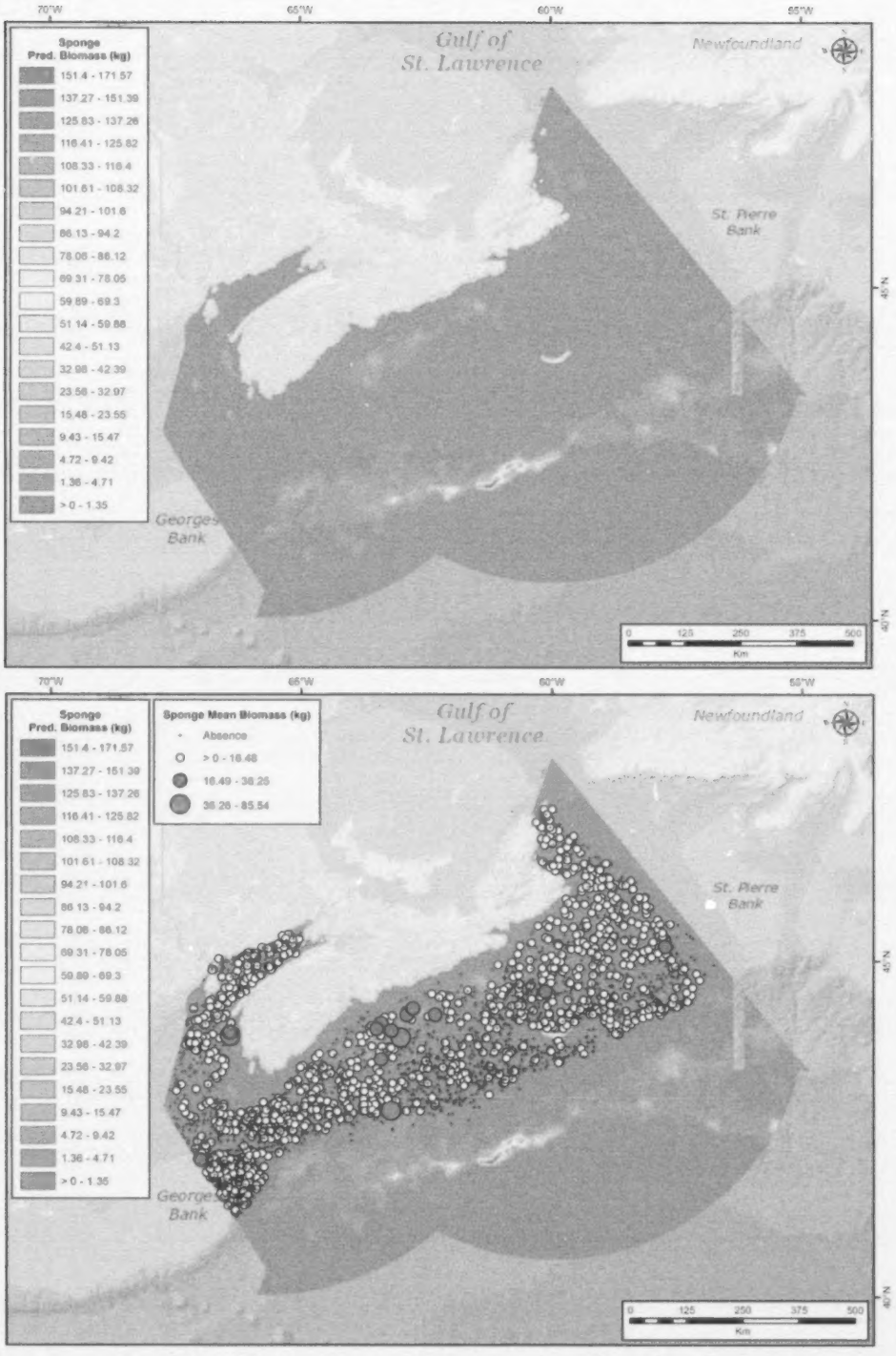

Figure A1.2. Prediction of sponge biomass (kg) from the GAM RF Variables model in the Maritimes Region. Bottom map shows the sponge mean biomass observations overlain.

Vazella pourtalesi (Russian Hat sponge)

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting the biomass distribution of the glass sponge *Vazella pourtalesi* are presented in Table A1.4. The GAM 0.7 Variables model performed better than the GAM RF Variables model, as indicated by the higher R^2, deviance explained, and AIC. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.5 and A1.6, respectively.

Figure A1.3 shows the graphical diagnostics for both models. Both showed fairly normal residuals. The residuals vs. linear predictor plot for the GAM 0.7 Variables model showed patterns indicative of heteroskedasticity. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models, particularly for higher catch values.

Figures A1.4 and A1.5 show the predicted biomass surface of *Vazella pourtalesi* generated from the GAM RF Variables and GAM 0.7 Variables models, respectively. For the GAM RF Variables model, the majority of the study extent was predicted to have low (> 0 – 0.67 kg) *V. pourtalesi* biomass. The highest predicted biomass value in this model was 85.06 kg, which is consistent with than the maximum observed catch (85.54 kg). High biomass was predicted to occur in Emerald Basin which is consistent with the random forest model on the same data (see Figures 41 and 42). The highest predicted biomass was located slightly northeast of the location of the highest mean catch of sponges. The GAM 0.7 Variables model predicted similar results (Figure A1.5), although the areas of high biomass were not as intense as the former model. The highest predicted biomass value in this model was 166.85 kg, which is higher than the maximum observed catch.

Table A1.4. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of *Vazella pourtalesi* in the Maritimes Region.

	GAM RF Variables	**GAM 0.7 Variables**
R^2	0.168	0.711
Deviance explained	80.00%	90.50%
AIC	2114.934	2097.098
BIC	2229.202	2343.160

Table A1.5. Results of the GAM RF Variables model built to predict the biomass of *Vazella pourtalesi* in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Salinity Average Maximum	1.169	1.726	0.098
Bottom Temperature Mean	1.931	6.544	4.270×10^{-4}*
Bottom Temperature Average Minimum	4.707	5.960	1.110×10^{-5}*
Fall Chlorophyll a Minimum	1.561	17.441	8.440×10^{-7}*
Fall Primary Production Average Range	2.986×10^{-4}	0.077	0.994
Summer Primary Production Mean	8.808×10^{-5}	0.029	0.998
Summer Primary Production Average Range	3.546	6.892	9.000×10^{-6}*
Surface Salinity Average Maximum	1.249	8.949	4.180×10^{-4}*

Table A1.6. Results of the GAM 0.7 Variables model built to predict the biomass of *Vazella pourtalesi* in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Minimum	4.501	3.684	0.002*
Bottom Salinity Average Minimum	7.121×10^{-5}	0.082	0.997
Bottom Temperature Average Minimum	4.864	6.418	2.440×10^{-6}*
Annual Chlorophyll *a* Range	5.192	5.175	2.030×10^{-5}*
Spring Chlorophyll *a* Maximum	6.564×10^{-1}	1.709	0.215
Depth	2.810	4.449	0.002*
Annual Primary Production Mean	2.111	5.377	0.001*
Annual Primary Production Average Minimum	3.186	3.499	0.007*
Fall Primary Production Average Maximum	1.133	3.739	0.035*
Fall Primary Production Average Minimum	5.584×10^{-1}	0.942	0.375
Spring Primary Production Average Maximum	1.732	13.055	1.160×10^{-6}*
Spring Primary Production Average Minimum	3.729	5.148	1.710×10^{-4}*
Surface Salinity Average Range	2.992	1.196	0.190
Slope	1.711×10^{-3}	0.208	0.979

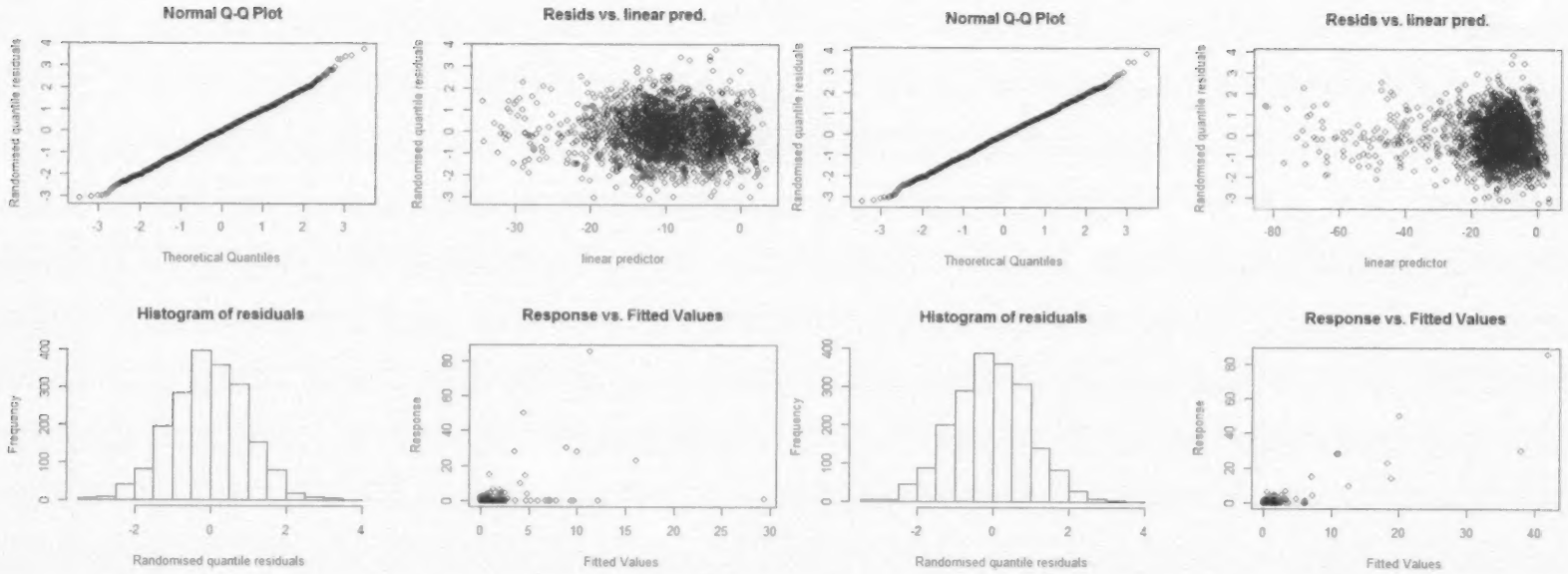

Figure A1.3. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the biomass of *Vazella pourtalesi* in the Maritimes Region.

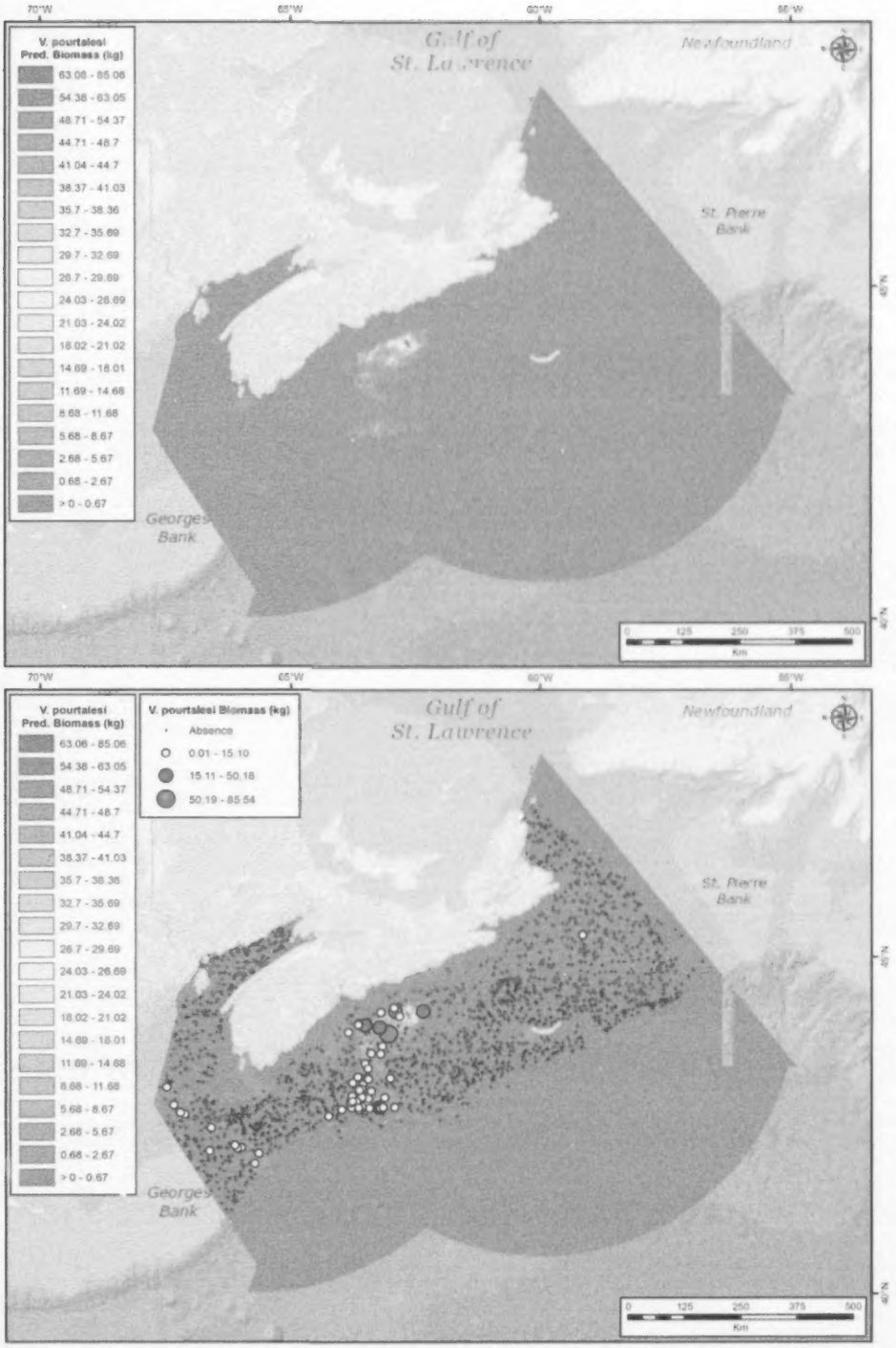

Figure A1.4. Prediction of *Vazella pourtalesi* biomass (kg) from the GAM RF Variables model in the Maritimes Region. Bottom map shows the *V. pourtalesi* mean biomass observations overlain.

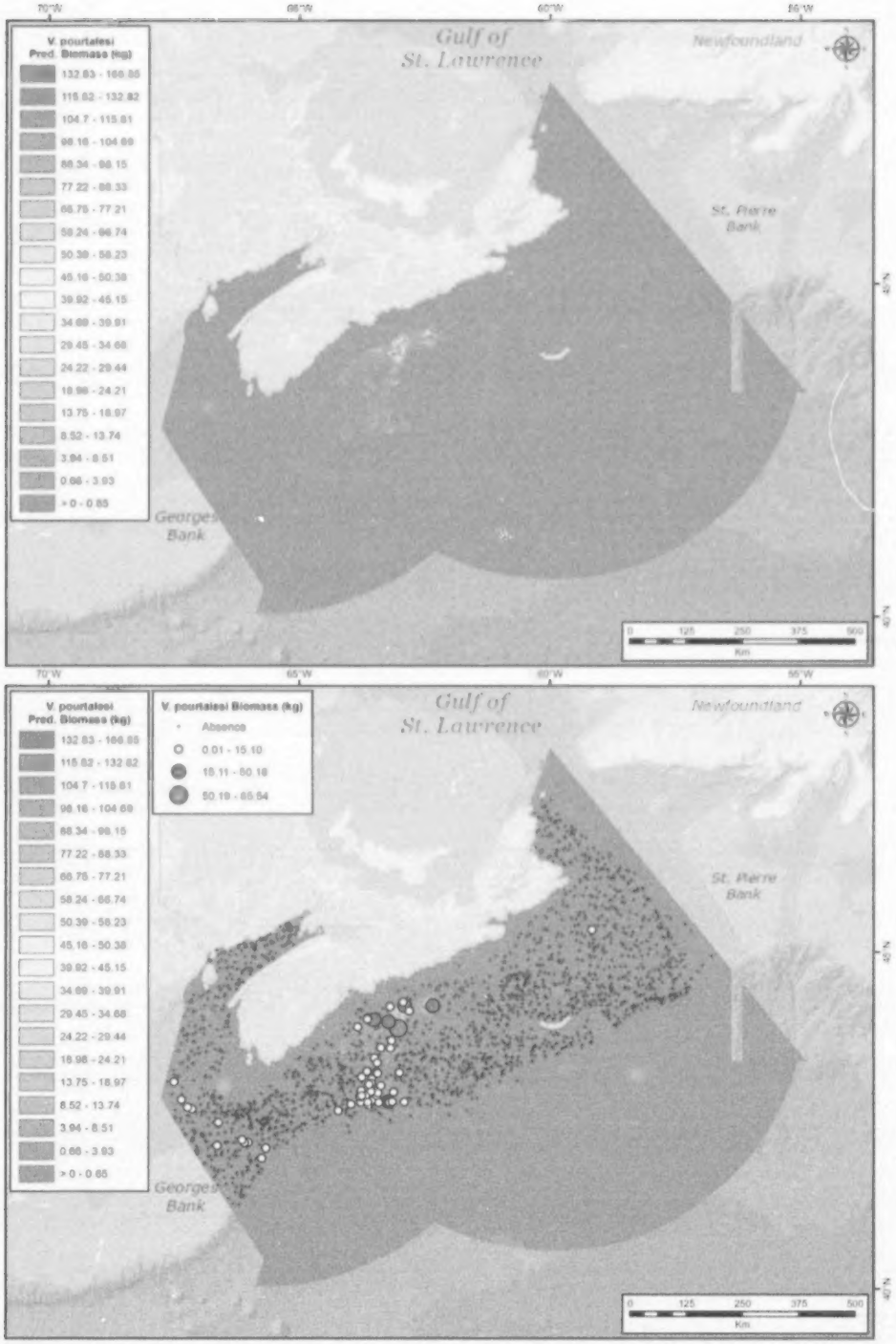

Figure A1.5. Prediction of *Vazella pourtalesi* biomass (kg) from the GAM 0.7 Variables model in the Maritimes Region. Bottom map shows the *V. pourtalesi* mean biomass observations overlain.

Sea Pens (Pennatulacea)

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean sea pen biomass are presented in Table A1.7. Both models performed fairly, with R^2 values of 0.268 and 0.247 for the GAM RF Variables and GAM 0.7 Variables models, respectively. The GAM RF Variables model performed better in terms of deviance explained and AIC and BIC. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.8 and A1.9, respectively.

Figure A1.6 shows the graphical diagnostics for both models. Both showed fairly normal residuals. The residuals vs. linear predictor plot for the GAM RF Variables model showed patterns indicative of heteroskedasticity. The response vs. fitted values plots showed a poor fit between the predicted and actual values for both models, particularly for higher catches.

When predicted to the entire extent of the study area, the models showed erroneously high predicted biomass values. High predictions of biomass were not alleviated with the inclusion of latitude and longitude or with modifications to the k value for each predictor. Predicted surfaces are therefore not presented for this taxonomic group.

Table A1.7. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of sea pens in the Maritimes Region.

	GAM RF Variables	GAM 0.7 Variables
R^2	0.268	0.247
Deviance explained	79.60%	73.80%
AIC	2273.103	2336.098
BIC	2615.161	2624.398

Table A1.8. Results of the GAM RF Variables model built to predict the biomass of sea pens in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Range	5.261	2.995	1.360×10^{-7}*
Bottom Salinity Average Maximum	7.362	7.978	5.540×10^{-16}*
Bottom Salinity Mean	7.697	7.311	2.190×10^{-14}*
Bottom Salinity Average Range	3.870	2.373	9.160×10^{-6}*
Bottom Shear Average Maximum	1.843	1.145	1.860×10^{-4}*
Bottom Shear Average Range	3.173	1.955	4.080×10^{-7}*
Bottom Temperature Average Range	5.922	5.437	3.160×10^{-11}*
Fall Chlorophyll a Mean	1.773	1.224	7.910×10^{-4}*
Depth	5.086	3.334	2.150×10^{-6}*
Surface Temperature Average Maximum	3.783	3.954	4.360×10^{-9}*

Table A1.9. Results of the GAM 0.7 Variables model built to predict the biomass of sea pens in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Minimum	1.370	17.367	6.670×10^{-7}*
Bottom Salinity Average Minimum	2.521	11.739	6.400×10^{-7}*
Bottom Temperature Average Minimum	3.483	8.043	9.710×10^{-7}*
Annual Chlorophyll a Range	0.425	0.865	0.441
Spring Chlorophyll a Maximum	3.087	2.135	0.055
Depth	1.586	5.609	0.002*
Annual Primary Production Mean	5.867	6.676	1.060×10^{-7}*
Annual Primary Production Average Minimum	3.713	8.418	4.120×10^{-7}*
Fall Primary Production Average Maximum	5.694	6.277	3.840×10^{-7}*
Fall Primary Production Average Minimum	4.217	5.417	5.160×10^{-5}*
Spring Primary Production Average Maximum	1.059	2.856	0.072
Spring Primary Production Average Minimum	0.599	1.240	0.302
Surface Salinity Average Range	1.940	9.920	2.220×10^{-5}*
Slope	1.325	4.573	0.017*

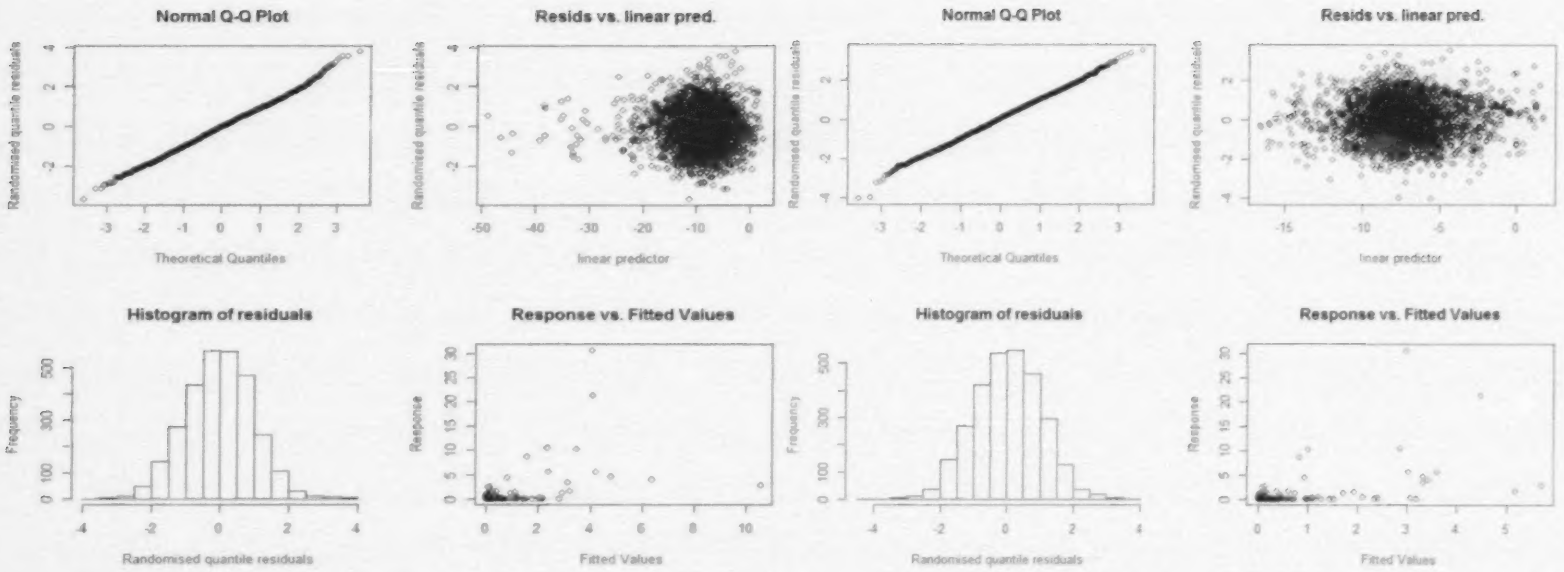

Figure A1.6. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the biomass of sea pens in the Maritimes Region.

Large Gorgonian Corals

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean large gorgonian coral biomass are presented in Table A1.10. Both models performed well, with R^2 values of 0.430 and 0.432 for the GAM RF Variable and GAM 0.7 Variable models, respectively. Deviance explained was higher for the GAM RF Variable model, and the AIC/BIC values were lower. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.11 and A1.12, respectively.

Figure A1.7 shows the graphical diagnostics for both models. Both models showed fairly normal residuals and only small patterns in the residuals vs. linear predictor plots. The response vs. fitted values plots showed a poor fit between the predicted and actual values, particularly for smaller catches.

When predicted to the entire extent of the study area, the models showed erroneously high predicted biomass values. High predictions of biomass were not alleviated with the inclusion of latitude and longitude or with modifications to the k value for each predictor. Predicted surfaces are therefore not presented for this taxonomic group.

Table A1.10. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of large gorgonian corals in the Maritimes Region.

	GAM RF Variables	GAM 0.7 Variables
R^2	0.430	0.432
Deviance explained	81.30%	71.80%
AIC	2207.499	2249.884
BIC	2351.450	2373.409

Table A1.11. Results of the GAM RF Variables model built to predict the biomass of large gorgonian corals in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Salinity Average Maximum	1.920×10^{-3}	0.236	0.977
Bottom Salinity Mean	6.841	10.201	7.580×10^{-14}*
Bottom Temperature Average Minimum	2.703	3.838	0.007*
Depth	4.018×10^{-1}	0.214	0.705
Spring Mixed Layer Depth Average Maximum	2.155	22.722	1.170×10^{-11}*
Fall Primary Production Average Minimum	9.668×10^{-1}	3.670	0.053
Surface Current Average Maximum	9.717×10^{-5}	0.094	0.996
Surface Current Mean	3.061×10^{-5}	0.009	0.999
Surface Current Average Range	2.789	10.267	2.770×10^{-7}
Slope	2.576	16.833	1.360×10^{-10}

Table A1.12. Results of the GAM 0.7 Variables model built to predict the biomass of large gorgonian corals in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Minimum	1.222	8.376	0.003*
Bottom Salinity Average Minimum	1.257	10.444	3.430×10^{-4}*
Bottom Temperature Average Minimum	3.224	5.787	1.170×10^{-4}*
Annual Chlorophyll *a* Range	1.715	6.490	0.001*
Spring Chlorophyll *a* Maximum	5.656×10^{-6}	0.001	0.999
Depth	5.398×10^{-5}	0.093	0.997
Annual Primary Production Mean	1.972×10^{-4}	0.068	0.996
Annual Primary Production Average Minimum	2.218	9.647	1.080×10^{-5}
Fall Primary Production Average Maximum	1.250	5.599	0.012*
Fall Primary Production Average Minimum	2.340×10^{-4}	0.009	0.998
Spring Primary Production Average Maximum	3.804×10^{-1}	0.701	0.514
Spring Primary Production Average Minimum	2.453×10^{-5}	0.088	0.998
Surface Salinity Average Range	8.899×10^{-1}	2.022	0.158
Slope	2.586	21.298	2.820×10^{-13}*

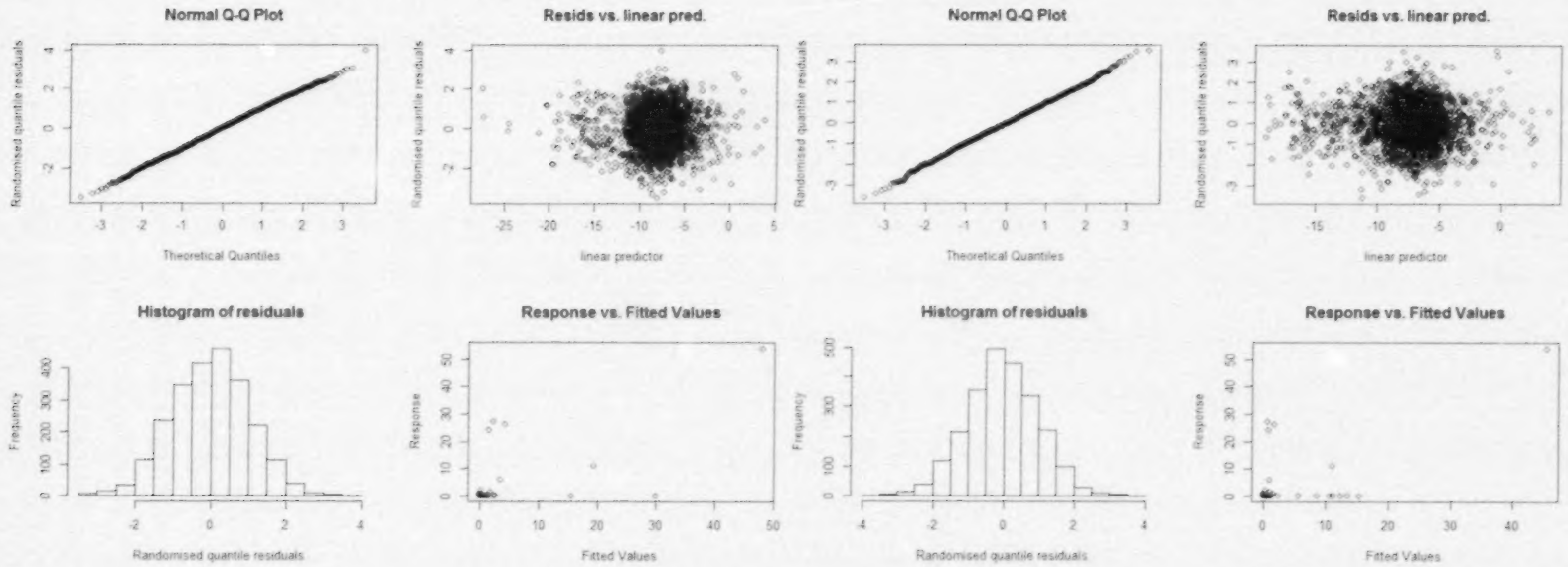

Figure A1.7. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the biomass of large gorgonian corals in the Maritimes Region.

Small Gorgonian Corals

The performance measures for both the GAM RF Variables and GAM 0.7 Variables models predicting mean small gorgonian coral biomass are presented in Table A1.13. Both models performed fairly, with R^2 values of 0.144 and 0.174 for the GAM RF Variable and GAM 0.7 Variable models, respectively. Deviance explained was higher for the GAM 0.7 Variable model. The variable significance for the GAM RF Variable and GAM 0.7 Variable models are shown in Tables A1.14 and A1.15, respectively.

Figure A1.8 shows the graphical diagnostics for both models. The residuals vs. linear predictor plots showed patterns indicative of heteroskedasticity, while the response vs. fitted values plots showed a poor fit between the predicted and actual values for both models.

Figures A1.8 and A1.9 show the predicted biomass surface of small gorgonian corals generated from the GAM RF Variables and GAM 0.7 Variables models, respectively. For the GAM RF Variables model, the majority of the study extent was predicted to have low (> 0 – 0.004 kg) small gorgonian biomass. Biomass was predicted to be high in the Laurentian Channel and along a narrow band on the Scotian Slope. The area of high biomass along the Scotian slope was associated with a cluster of large biomass values, and is consistent with the random forest model results (see Figures 109 and 110). The GAM 0.7 Variables model predicted high biomass of small gorgonian corals in the same area (Figure A1.9). The highest predicted biomass value in this model was 0.46 kg, which is only slightly higher than the maximum mean biomass catch in the raw data (0.34 kg).

Table A1.13. Comparison of the accuracy measures between the GAM RF Variables and GAM 0.7 Variables models built to predict the biomass of small gorgonian corals in the Maritimes Region.

	GAM RF Variables	GAM 0.7 Variables
R^2	0.144	0.174
Deviance explained	53.40%	71.50%
AIC	1633.629	1638.405
BIC	1696.464	1772.240

Table A1.14. Results of the GAM RF Variables model built to predict the biomass of small gorgonian corals in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Minimum	3.397×10^{-1}	0.817	0.501
Bottom Salinity Average Minimum	6.149×10^{-1}	0.922	0.355
Bottom Salinity Average Range	1.396	7.038	0.001*
Bottom Shear Average Minimum	3.351×10^{-5}	0.500	0.995
Summer Chlorophyll a Mean	1.465	8.081	4.00×10^{-4}*
Summer Chlorophyll a Minimum	3.494×10^{-6}	0.171	0.999
Depth	2.894	6.354	2.300×10^{-4}*
Spring Primary Production Average Maximum	6.195	0.272	0.995
Surface Temperature Average Maximum	1.478×10^{-5}	0.179	0.998
Surface Temperature Mean	1.346×10^{-5}	0.136	0.999

Table A1.15. Results of the GAM 0.7 Variables model built to predict the biomass of small gorgonian corals in the Maritimes Region. The estimated degrees of freedom (edf), *F* value, and *p*-value are shown for each variable. Significance was tested at the $\alpha = 0.05$ level. Significant variables are indicated with an asterisk (*).

Variable	edf	*F*	*p*-value
Bottom Current Average Minimum	4.007×10^{-6}	0.028	1.000
Bottom Salinity Average Minimum	5.431	3.656	0.001*
Bottom Temperature Average Minimum	1.922	6.852	3.930×10^{-4}*
Annual Chlorophyll *a* Range	5.148×10^{-1}	0.788	0.421
Spring Chlorophyll *a* Maximum	2.581×10^{-5}	0.007	1.000
Depth	2.682	5.679	7.470×10^{-4}*
Annual Primary Production Mean	1.528×10^{-6}	0.005	1.000
Annual Primary Production Average Minimum	2.572	1.527	0.209
Fall Primary Production Average Maximum	1.410×10^{-1}	0.552	0.708
Fall Primary Production Average Minimum	1.211	4.830	0.011*
Spring Primary Production Average Maximum	4.884×10^{-6}	0.031	1.000
Spring Primary Production Average Minimum	2.684	7.295	4.940×10^{-5}*
Surface Salinity Average Range	9.713×10^{-7}	0.002	1.000
Slope	1.680×10^{-5}	0.013	0.999

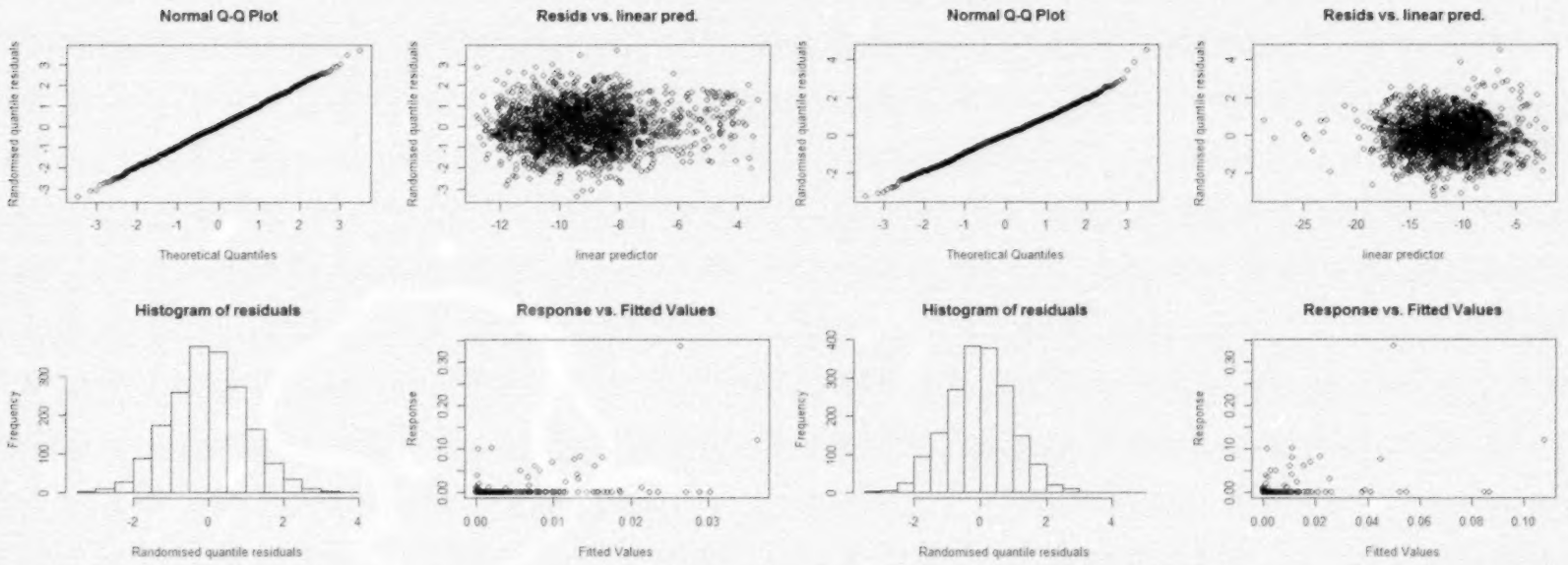

Figure A1.8. Model diagnostics for the GAM RF Variables model (left) and the GAM 0.7 Variables model (right) built to predict the biomass of small gorgonian corals in the Maritimes Region.

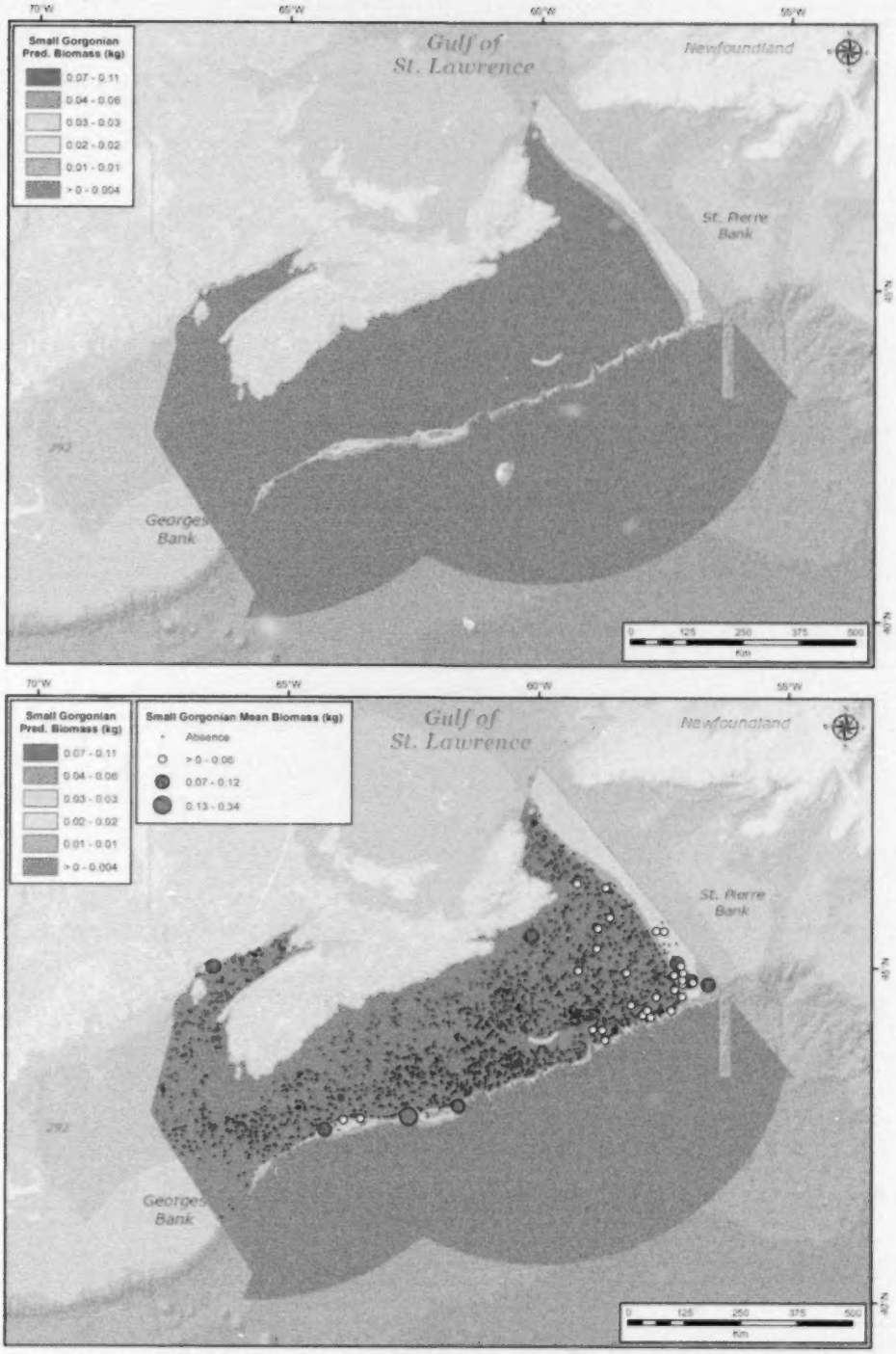

Figure A1.9. Prediction of small gorgonian coral biomass (kg) from the GAM RF Variables model in the Maritimes Region. Bottom map shows the small gorgonian coral mean biomass observations overlain.

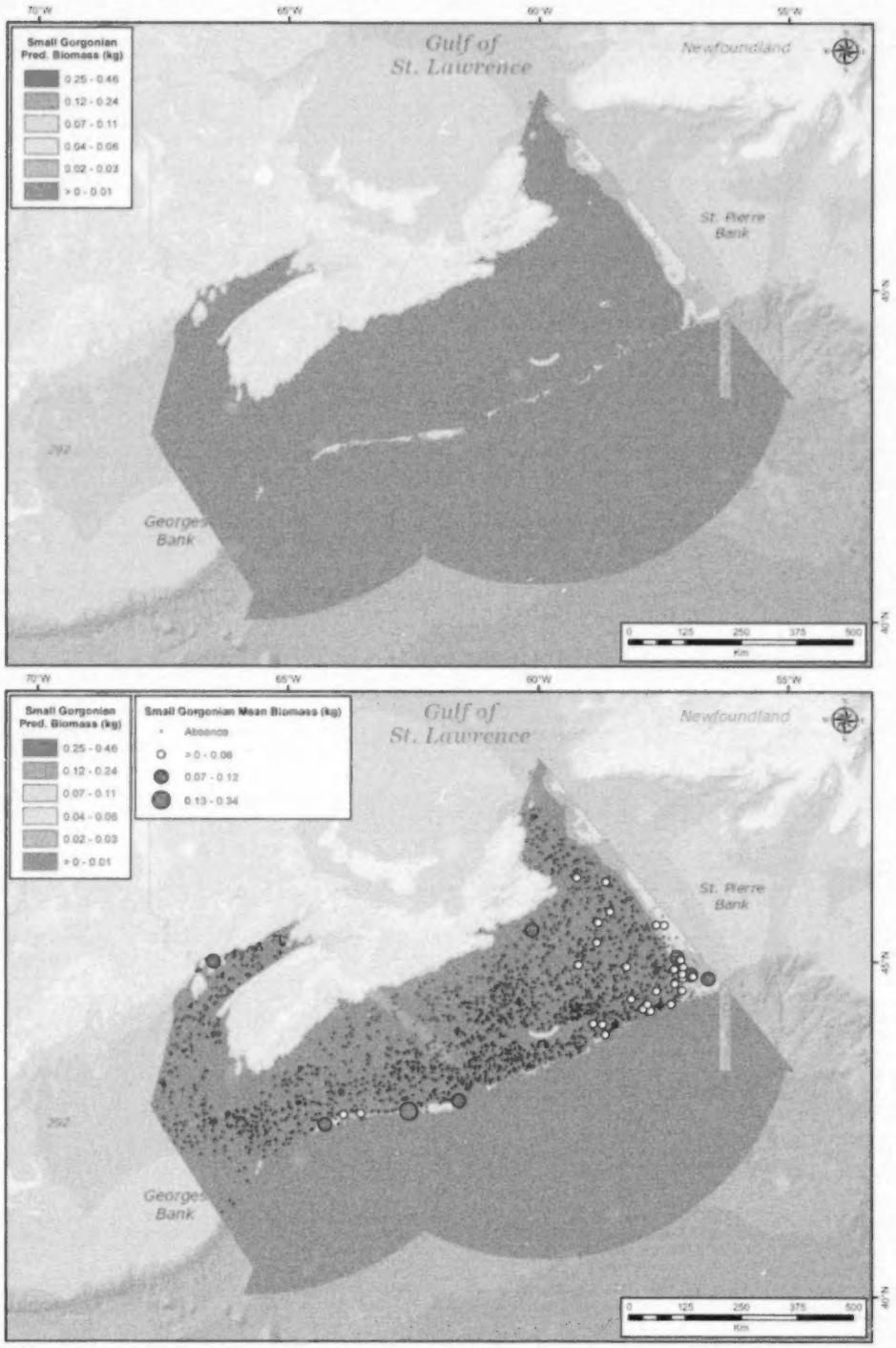

Figure A1.10. Prediction of small gorgonian coral biomass (kg) from the GAM 0.7 Variables model in the Maritimes Region. Bottom map shows the small gorgonian coral mean biomass observations overlain.